A computer-aided mass diagnosis system based on perceptive features learned from quantitative mammography radiology report: An observer-based study

Zilong He
Department of Radiology, Nanfang Hospital, Southern Medical University, Guangzhou
Yue Li
School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou
Weimin Xu
Department of Radiology, Nanfang Hospital, Southern Medical University, Guangzhou
Chanjuan Wen
Department of Radiology, Nanfang Hospital, Southern Medical University, Guangzhou
Xiangyuan Ma
Department of Biomedical Engineering, College of Engineering, Shantou University,Shantou
Jun Wei
Perception Vision Medical Technologies LTD. Co, Guangzhou
Hui Zeng
Department of Radiology, Nanfang Hospital, Southern Medical University, Guangzhou
Zeyuan Xu
Department of Radiology, Nanfang Hospital, Southern Medical University, Guangzhou
Sina Wang
Department of Radiology, Nanfang Hospital, Southern Medical University, Guangzhou
Jiefang Wu
Department of Radiology, Nanfang Hospital, Southern Medical University, Guangzhou
Chenya Feng
Department of Radiology, Nanfang Hospital, Southern Medical University, Guangzhou
Mengwei Ma
Department of Radiology, Nanfang Hospital, Southern Medical University, Guangzhou
Genggeng Qin
Department of Radiology, Nanfang Hospital, Southern Medical University, Guangzhou
Yao Lu
School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou
Weiguo Chen (chen1999@smu.edu.cn)
Department of Radiology, Nanfang Hospital, Southern Medical University, Guangzhou

Research Article

Keywords: Computer-aided diagnosis, Digital mammographic, Convolutional neural network, Mass

Posted Date: February 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-222002/v1

Title page

A computer-aided mass diagnosis system based on perceptive features learned from quantitative mammography radiology report: An observer-based study

Zilong He[1,*], Yue Li[2,*], Weimin Xu[1], Chanjuan Wen[1], Xiangyuan Ma[2,5], Jun Wei[3], Hui Zeng[1], Zeyuan Xu[1], Sina Wang[1], Jiefang Wu[1], Chenya Feng[1] Mengwei Ma[1], Genggeng Qin[1], Yao Lu[2, 4] and Weiguo Chen[1]

[1] Department of Radiology, Nanfang Hospital, Southern Medical University, Guangzhou, China
[2] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[3] Perception Vision Medical Technologies LTD. Co, Guangzhou, China
[4] Guangdong Province Key Laboratory Computational Science, Sun Yat-Sen University, Guangzhou, China
[5]Department of Biomedical Engineering, College of Engineering, Shantou University.
[*] Authors contributed equally to this work.

Corresponding authors:
Weiguo Chen:
Email: chen1999@smu.edu.cn

Yao Lu:
Email: luyao23@mail.sysu.edu.cn

Genggeng Qin:
Email: zealotq@smu.edu.cn

Abstract

Background: Computer-aided diagnosis (CAD) system can provide reference to radiologists in breast mass classification. This study was to verify if a CAD model, based on perceptive features learned from quantitative BI-RADS descriptions, can help radiologists improve diagnosis performance for breast masses in mammography.

Methods : A retrospective multi-reader multi-case (MRMC) study was conducted to evaluate a CAD model established on perceptive features. Digital mammograms of 416 patients with breast masses were collected from 2014 to 2017, including 231 benign and 185 malignant masses. Altogether, 214 of them (109 benign, 105 malignant) were selected randomly to train the CAD model which consisted of perceptive feature extractor and classifier. The other 202 patients were used as the test set for evaluation from which 51 patients (29 benign and 22 malignant) were selected. Six radiologists were divided into three groups (junior, middle-senior, and senior).They evaluated 51 patients without and with support from the CAD model. BI-RADS category, benign or malignant diagnosis, probability of malignancy, and diagnosis time were recorded during two evaluation sessions.

Results: In the MRMC evaluation, the average AUC of six radiologists with CAD support was significantly higher than that without support (0.896 vs. 0.850, p=0.02). Both of average sensitivity and average specificity increased (p=0.0253). More cases were assessed as BI-RADS 4 than BI-RADS 2 or 3. Five radiologists showed comparable diagnosis time per case with and without CAD support, and one radiologist showed a significant decrease when the CAD model was involved.

Conclusion: The CAD model could improve radiologists' diagnostic performance for breast masses without improving the diagnosis time.

Keywords: Computer-aided diagnosis; Digital mammographic; Convolutional neural network; Mass

1. Background

Full-field digital mammography (FFDM) is considered as an effective method for breast cancer screening [1,2]. In developed and developing countries, it has become the first option for routine medical exam [3]. However, the growing number of women screening examinations has resulted in an increasing workload of radiologists. The overall diagnosis time of each patient showed an upward trend, indicating that the work efficiency of radiologists decreased [4]. Lu S et al. showed that based on China's huge population, the increasing workloads of radiologists was accompanied by a decline in work efficiency [5]. In addition, Karssemeijer et al. illustrated a positive correlation between the increased workload of radiologists and the demand for breast screening [6].

Due to the lack of experienced radiologists, junior radiologists participated in screening prematurely without training, resulting in a decline in diagnostic accuracy and sensitivity of breast cancer and increased risk of misdiagnosis and missed diagnosis [7,8,9]. As the glandular type in Chinese women is generally dense breast, it further increases the difficulty for junior radiologists to recognize the characteristics of breast cancer, especially the margins and shape of breast cancer with mass as the main sign [10,11,12]. Friedewald et al. illustrated that radiologists, especially inexperienced junior ones, had decreased diagnostic sensitivity in dense breast [13]. Broeders et al. believed that junior radiologists were inexperienced in the

characteristics of the mass of breast cancer, which led to inaccuracy in BI-RADS category evaluation and affected the prognosis of patients [14].

Computer-aided diagnosis (CAD) systems were introduced as auxiliary methods to improve radiologists' diagnostic efficiency. For the breast mass classification task, feature extraction is an important step. Conventional CAD methods extract several hand-crafted features from the region of interest (ROI) to form the feature vector for each mass, which are of three types – intensity, shape, and texture. [15]. In addition, deep learning technology has recently been used for feature design. Deep learning model can learn the latent features directly from the ground truth so that more represented features can be designed [16,17,18]. For example, Jiao et al. used convolution neural network (CNN) pre-trained on ImageNet as a feature extractor for the breast mass in breast cancer diagnosis [19]. Kooi et al. combined the deep features and conventional hand-crafted features to distinguish the true mass from normal breast mammary tissue. Their results showed that the combined feature set performed best in the classification stage [20].

Hand-crafted features are designed based on human's experience, but they are not task-specific so that they may perform not so well in medical imaging analysis. Deep learning model learns features by optimizing weights according to our object of task, but this procedure lacks human's experience as reference, which is important in clinic. In order to combine radiologists' clinical experience and deep learning method together to design proper features for mass diagnosis, this study proposed a training scheme that took BI-RADS descriptions as into consideration. These features were referred to as perceptive features and a CAD model was established based on them. An observer study was also conducted to evaluate the use of this model in assisting radiologists in diagnosis. The results showed that that radiologists, especially junior ones, significantly improved their performances with the support of our proposed model while not increasing their workloads.

2. Materials and methods

To ensure that a deep learning-based model could obtain a sufficient quantitative ability, we used the BI-RADS characteristic description of a mass, to train a CNN as a perceptive feature extractor. To realize this goal, description quantification, feature extractor training, and classifier training were needed. An observer study was then conducted to verify the clinical meaning of this model. These steps are illustrated in detail as follows.

2.1 Dataset and mammograms collection

We retrospectively retrieved samples between April 2014 and October 2017 from Nanfang Hospital of Southern Medical University, Guangzhou, China. Since this study desired to establish a CAD model for breast masses' benign and malignant classification, each case collected in this study only had one mass in unilateral breast. Pathological result of each mass was used as classification ground-truth. Cases with calcification, architectural distortion, or asymmetries were not considered in this study in order to avoid confused ground-truth because each case only had one biopsy result. There was no positive sign in contralateral breast (BI-RADS category 1 or 2). In addition, all collected cases had bilateral craniocaudal (CC) and mediolateral oblique (MLO) images, clinical medical history, radiology reports, and operative and pathological findings. Women with implants, lesions not fully visible, or large lesions occupying almost the whole breast in CC or/and MLO mammograms were excluded.

In total, 416 cases that met the above inclusion criteria were obtained. Of these, 214 were chosen randomly as training sets, which were used to train the feature extractor and mass classification model. The remaining 202 cases were used as independent test sets to evaluate the model, out of which, 51 cases were randomly selected for the observer study. All of them were anonymized and represented by a new ID.

Table 1 Characteristics of the population for this study.

Variable	Training Set (n=214)	Test Set (n=202)	51 Cases in Observer Evaluation
Patient age (y)			
Mean	45.64	45.51	46.53
Median	45	45	47
Range	23-73	23-78	27-65
Interquartile Range	40-50	40-50	40-51
BI-RADS breast density			
a	6	5	3
b	23	25	9
c	169	155	35
d	16	17	4
Number of each class			
Benign	109	122	29
Malignant	105	80	22

The characteristics of the population and digital mammograms are shown in Table 1. The digital mammography was performed using the Selenia Dimensions System, Hologic, Bedford, MA, USA. The size of each image was 3328×2560 with pixel spacing of 0.06mm.

Other detailed information about these 416 cases is provided in Appendix A.

2.3 The region of interest selection

The ROIs of all masses were marked by an experienced radiologist (X. Liao, with 15 years' experience in digital mammography) by delineating the masses' boundary. Three radiologists (GG. Qin, L. Zhang, and WG Chen, also with 15 years' experience) reviewed the ROI, and all three did not take part in the observer study. If there was any disagreement about the location or shape of a certain mass among these three radiologists, they would determine a final ROI by voting.

2.4 Mass classification model

2.4.1 Quantification of the BI-RADS description

We hoped that our CAD model could extract perceptive features that reflect semantic characteristics such as human's vision perception and diagnosis experience. This experience is reflected in radiology reports. Several descriptions defined in BI-RADS lexicon for mass include shape, margins, and density, which are the main factors for radiologists to diagnose breast cancer.

Different descriptions have different probabilities of malignant masses. For example, irregular shapes or circumscribed margin are correlated with suspicious findings, and oval shapes or indistinct margin are correlated with benign findings. The method to employ these descriptions in perceptive feature design is to use them as the ground-truth to train a regression network. The feature vector in the last fully connected layer can be regarded as the perceptive feature that we want.

Table 2 Specific quantification for each description.

Descriptions	Radiologists' Assessment	Quantification
Shape	Oval or Round	0
	Irregular	1
Margins Sharpness	Circumscribed	0
	Obscured	0.5
	Indistinct	1
Microlobulated Margins	No	0
	Yes	1
Spiculated Margins	No	0
	Yes	1
Density	Low or Fat-containing	0
	Equal	0.5
	High	1

However, training procedure requires quantitative ground-truth instead of text descriptions. It is necessary to quantify these descriptions. To correlate the quantification to the classification task, we quantify descriptions as a malignancy probability. Descriptions about malignant, uncertain and benign findings are quantified as 1, 0.5 and 0 respectively. Details for quantification is shown in Table 2 and an example is shown in Figure 1.

2.4.2 Stage 1: Feature Extractor

The backbone of the feature extractor is the classical CNN VGG16[21]. It is used to classify the class of objects in natural images and plays an important role in computer-aided diagnosis. In this study, we made a slight modification to VGG16 in order to meet our needs.

A 2-channel patch with size of 288×288 centered at a mass is the input of network. One channel is original FFDM and another is binary mask represents the ROI. The remaining convolution layers, activation functions and pooling layers are the same as the original VGG16 network. Then, three fully connected layers with Rectified Linear Unit (ReLU) activation functions and dropout operations are used to convert this feature map into a feature vector. Finally, the network outputs a 5-dimension vector, which represents the predicted quantitative descriptions of the input mass. The specific architecture of this feature extractor is shown in Figure 2.

Mean square error is used as loss function. The weights of each layer were initialized randomly according to a standard normal distribution and updated by an Adam optimizer during the training process. Both CC-view and MLO-view masses were fed into the same feature extractor. In the training process, the masses of the same patient in CC-view and MLO-view FFDM, shared the same quantitative BI-RADS descriptions.

After the VGG16 network was trained, we discarded its output layer so that the remaining network will output a 128-dimension vector, which is the perceptive feature vector we need.

2.4.3. Stage 2: Benign and malignant classification

These features are then used to train a classifier that can distinguish between benign and malignant masses. Stepwise regression feature selection method and linear discriminant analysis (LDA) classifier were employed to realize this goal.

In the training of stepwise regression and LDA, we did not differentiate between masses from CC-view and MLO-view images, that is, a lesion-wise classification model was

considered. In the test process, the malignancy probability output by model from CC-view and MLO-view images of the same case were averaged for case-wise evaluation.

2.4.4 Model Selection and Test

Ten-fold cross validation was used. The 214 training cases were randomly divided into ten folds. In each training time, 9 folds were used as the training set and 1 fold was used as the validation set. Feature extractor was trained until the loss of both the sets plateaued. After ten repetitions, all folds were used as the validation set once and ten trained models were obtained.

During the test process, ten trained models were used on the 202 independent test cases and output ten predicted scores for each case.

Models fusion always obtains a better model whose performance beyond each individual one. To fuse these ten trained models, the averaged probability of malignancy (POM) among these ten models was calculated for each case, which was used in the multi-reader multi-case (MRMC) evaluation and stand-alone study.

2.5 Observer evaluation

2.5.1 MRMC evaluation

Evaluation was separated by two sessions, one without our model's results as reference and one with this reference. Both sessions were performed on 51 cases. The time interval between the two sessions was more than 15 days.

To avoid individual diagnosis difference in different time for a same radiologist on same cases, more than one radiologist participated in this observer study. In total, six radiologists performed this evaluation. Two of them were junior radiologists with 2- year experience (reader 1: CY. Feng and reader 2: MW. MA), two of them were middle-seniority radiologists with 4-year experience (reader 3: ZY XU and reader 4: SN. Wang) and the remaining two radiologists were senior with 6-year experience (reader 5: JF. Wu and reader 6: H. Zeng).

In the first session, each radiologist observed the FFDM of the cases without our proposed model support (unaided evaluation). Only the original images and ROIs of each mass in both CC-view and MLO-view FFDM were provided. In the second session, in addition to the FFDMs and ROIs of masses, the POM calculated by our proposed classification model was also provided to radiologists (aided evaluation). In both of these two sessions, the BI-RADS category, benign or malignant classification, POM (ranged from 0-100%), consuming time, and so on were recorded by observers for statistical analysis. These results are recorded in an example table in Appendix B.

Before these two sessions, six radiologists learnt evaluation criteria and announcements. They would be informed that each case only had one lesion. History-taking, results of other examinations, and palpation were blinded for them. 20 example cases, were not included in the 51 observation cases, were used to train radiologists performing this process.

2.5.2 Stand-alone evaluation

The performance of the classification model for 202 independent cases was also compared with two senior experienced radiologists (reader 7: WM. Xu and reader 8: CJ. Wen) who did not participate in the MRMC evaluation. Each radiologist read the FFDM of 202 independent test cases without any reference and gave a POM for each case. The ROC performances of radiologists and the classification model were compared to show the difference between humans and our model.

2.6 Statistical analysis

In MRMC study, the area under receiver operating characteristic (ROC) curve (AUC) was calculated according to POM assessed by radiologists for each session. ROC curves were

obtained by ranking all POMs evaluated by certain radiologist on certain case set in ascending order. True positive rate (TPR) and false positive rate (FPR) were calculated at the probability threshold of each ranked POM. Taking all of the TPRs as coordinates in y-axis and all of the FPRs as coordinates in x-axis will draw the ROC curve for this radiologist on this case set.

To analyze the significance between two evaluation session, Wald or z-test was used to yield a p-value with the null hypothesis of that these two sessions had same AUCs. For comparison of sensitivity and specificity between two sessions, the assessment results of benign and malignant masses were compared with the biopsy-proven ground truth. Binary version MRMC analysis was implemented to yield a p-value. The average diagnosis time of each case was calculated for each radiologist on each session and paired t-test was used to yield p-value for the difference between two sessions.

In stand-alone study, the ROC curves and AUCs were used to compared the performances between senior radiologists and our proposed CAD model and p-value was calculated by ROC test.

Package in R language of 'iMRMC', 'ROCR' and 'pROC' were used to conduct statistical analysis in this study.

3. Results

3.2 Parameters Selection

The learning rate of the Adam optimizer was initialized as 0.0001. There were two decay times with decay gamma of 0.1 at epoch 30 and epoch 60, respectively. Stepwise regression and LDA classifier were implemented by MATLAB 2018a with default parameters.

3.3 ROC performance

In MRMC study, readers significantly improved their diagnosis performance with support of our proposed CAD model. The averaged AUCs increased from 0.850 to 0.896 ($p = 0.0209$). ROC curves changes are shown in Figure 3 and Specific AUC changes are shown in Table 3. Five of six radiologists' AUCs increased with reference to our proposed model and one radiologist's AUC decreased.

Table 3 The AUCs in multi-reader multi-case observer study.

Reader	AUC unaided	AUC with model reference	Difference	P value
1	0.842	0.920	0.078	
2	0.783	0.892	0.109	
3	0.889	0.922	0.033	
4	0.852	0.890	0.038	
5	0.866	0.904	0.038	
6	0.869	0.847	-0.022	
Diagonal Average	0.850	0.896	0.046	0.0209

3.4 Benign and malignant evaluation

The sensitivity and specificity for radiologists in the two sessions are shown in Table 4. All radiologists' sensitivities with model support were higher than or equal to that without model support. It was most obvious in junior-group radiologists. Five of six radiologists' specificities were higher than or equal to that without model support. Only reader-5's specificity decreased. Binary MRMC analysis showed the performance improvement was significant ($p = 0.0253$)

Table 4 The difference in sensitivity and specificity in three groups with and without model reference

Sensitivity:				
Group	**Reader**	**Sensitivity unaided**	**Sensitivity aided**	**Difference**
Junior	1	0.545	0.682	0.137
	2	0.682	0.864	0.182
Middle-seniority	3	0.773	0.773	0
	4	0.773	0.864	0.091
Senior	5	0.819	0.901	0.082
	6	0.773	0.773	0
Specificity:				
Group	**Reader**	**Specificity unaided**	**Specificity aided**	**Difference**
Junior	1	0.931	0.931	0
	2	0.621	0.621	0
Middle-seniority	3	0.793	0.897	0.104
	4	0.862	0.931	0.069
Senior	5	0.793	0.689	-0.104
	6	0.863	0.897	0.034

3.5 BI-RADS Evaluation

In MRMC evaluation, all the readers adjusted the BI-RADS category of partial cases with model support, which focused on BI-RADS 2,3,4. The assessments tended toward an increase in BI-RADS 4, and fewer cases were defined as BI-RADS 2 or 3 with the model reference. In total, 80 cases' BI-RADS assessments increased while 48 cases' BI-RADS assessments decreased. More details are shown in Appendix C.

3.6 Diagnosis time

Diagnosis time was recorded by each radiologist using timer software with number of seconds. The average diagnosis time per case for radiologists in these two sessions are shown in Table 5. Five of six radiologists had comparable diagnosis efficiency. Reader 6's diagnosis time significantly decreased from 56.96 to 43.96 after involving the CAD support ($p = 0.01$). Two senior experienced radiologists showed a larger decreasing diagnosis time than other radiologists. Figure 4 shows the reading time comparison for all readers and reader 6.

Table 5 The mean diagnosis time for radiologists in multi-reader multi-case study.

Reader	**Mean Time w/o support(s)**	**Mean Time with model support (s)**	**Difference**	**P-value**
1	55.27	55.51	0.24	0.955
2	80.59	81.18	0.59	0.912
3	63.90	64.24	0.34	0.928
4	45.10	42.59	-2.51	0.378
5	42.35	37.35	-5	0.089
6	56.96	43.96	-13	0.001

3.7 Stand-alone study

In the stand-alone study, the ROC curves in 202 independent cases of our proposed model and two senior experienced radiologists are shown in Figure 5. Our proposed model achieved AUC of 0.913. Reader 7 and reader 8 achieved AUCs of 0.969 and 0.988 respectively. Both radiologists' performances were better than our CAD model. This result showed that although this model improved the diagnosis efficiency and accuracy for radiologists with experience less than 6 years. It cannot attach the performance with experience more than 8 years.

4. Discussion and conclusions

In this study, we proposed a deep-learning-based perceptive feature extractor for breast mass in order to establish a CAD model for benign and malignant mass classification and evaluated this CAD model through observer study. The results showed that this CAD model could assist radiologists to improve diagnosis accuracy while not increased diagnosis time.

This perceptive feature extractor used quantitative BI-RADS descriptions instead of the biopsy-proven results to optimize its weights. This brought benefits to CAD model. First, BI-RADS descriptions were obtained from the radiologists. When optimizing, the human vision perception and clinical experience were integrated into the weights. It provided us more ideas to interpret the learned features of CNN. Second, compared to using a CNN directly to establish a CAD model, training the feature extractor first and then using a classifier to finish the diagnosis, divided this process into two stages, which was more consistent with the process of clinical diagnosis.

In the observer study, the AUCs showed that radiologists had a higher and better diagnostic performance with CAD model support than that without model support. In addition, the diagnostic sensitivities of all radiologists increased when the model involved [22]. Except for reader 5, other readers' specificities increased or remained the same after the model references were used. In particular, for the junior radiologists, their AUCs and sensitivities increased in a larger range than middle-seniority and senior radiologists while their specificities remained unchanged [23, 24, 25].

In addition, radiologists adjusted the evaluation of BI-RADS category after this CAD model involved. More BI-RADS 4 and less BI-RADS 2 or 3 were assessed, which means that the model could assist radiologists in observation, increasing their attention in the most suspicious characteristic.

Considering the workload of radiologists in daily work, we do not hope using CAD model will increase their diagnosis time for each patient. The average diagnosis showed that most radiologists had comparable diagnosis time and the other one radiologist had significant decreasing of diagnosis time [26]. This meant this model will not increase radiologists' workload.

A stand-alone study showed that the average AUC performance of the CAD model was weaker than that of senior radiologists with 8-year mammogram reading experience. This demonstrated that although this CAD model could help radiologists improve their diagnosis accuracy, it could not replace radiologists to make diagnosis.

Our study has some limitations. First, this is not an end-to-end CAD system. Automatic mass detection and segmentation of mass were not explored. To generate a classification probability, the radiologist must first locate the breast mass and mark the contour of the mass. Second, this is a single-center study. Different regions have different population distribution. We did not explore if this CAD model can be applied to other population outside our institution.

In the future, we will improve this model as an end-to-end CAD system. The model should detect and segment mass automatically so that radiologists can not only obtain the reference POM, but also the mass location and shape information, which may provide more useful tips. Since it increases the complexity of the whole task, we need more data to train a stable system.

we tend to collect more FFDM data from different data center with mass to address the limitation of limited data size and develop multi-center study.

List of abbreviations

FFDM: Full-field digital mammography
CAD: Computer-aided diagnosis
ROI: Region of interest
CC: Craniocaudal
MLO: Mediolateral oblique
ReLU: Rectified linear unit
MRMC: Multi-reader multi-case
POM: Probability of malignancy
CNN: Convolution neural network

Figure legends

Figure 1. An example of quantification for a malignant mass in MLO-view FFDM. Five text descriptions assessed by radiologist as shown in red box are quantified as corresponding numbers. A five-dimension vector is generated, which is used as the ground-truth to train the perceptive feature extractor.
Figure 2. The architecture of this feature extractor.
Figure 3. ROC curves for six readers in three groups diagnosis with and without CAD model support. (a)Junior group of reader 1 and 2; (b) Middle-seniority group of reader 3 and 4; (c) Senior group of reader 5 and 6.
Figure 4. Diagnosis time comparison. (a) The time comparison of all readers; (b) the time comparison for reader 6, who was the only one that showed obvious difference between two sessions. Graph shows differences in diagnosis time per case for all reader. Each red point indicated diagnosis time for a certain case with or without model support. There is no significant change when the point falls on diagonal. Point above the diagonal indicates diagnosis time has increase with model support. Point below the diagonal means the time decrease with model support.
Figure 5. The ROC curves of radiologists and our proposed model in 202 independent test cases.

Declarations

Ethics approval and consent to participate

This study has obtained ethics approval and consent from Medical Ethics Committee of Nanfang Hospital. The code number of clinical ethics project is NFEC-2018-037.

In this study, all methods were carried out in accordance with relevant guidelines and regulations, providing by the Medical Ethics Committee of Nanfang Hospital. And also, all participants' consent were waived by the Medical Ethics Committee of Nanfang Hospital.

Consent for publication

All the consent were waived by the Medical Ethics Committee of Nanfang Hospital.

Availability of data and materials

The datasets analyzed during the current study are not publicly available due to patient privacy information protection but are available from the corresponding author on reasonable request.

Competing interests

Y.Lu is founder and president of Perception Vision Medical Technologies LTD. Co. J.Wei is Chief Technology Officer and vice president of Perception Vision Medical Technologies LTD. Co. The remaining authors declare that they have no competing interests.

Funding

This study has received funding in part by National Key R&D Program of China under Grant Nos. 2018YFC1704206 and 2016YFB0200602, 2019YFC0121903 and 2019YFC0117301, the Fundamental Research Funds for the Central Universities under Grant No. 19LGYJS63, the NSFC under Grant Nos. 81971691, 81801809, 81830052, 81827802, U1811461, and 11401601, the Natural Science Foundation of Guangdong Province, China under Grant No. 2019A1515011168, and 2018A0303130215, the Science and Technology Innovative Project of Guangdong Province under Grant Nos. 2016B030307003, 2015B010110003, and 2015B020233008, the Science and Technology Planning Project of Guangdong Province under Key Grant Nos. 2015B020233002, and 2017B020210001, the Guangzhou Science and Technology Creative Project under Key Grant No. 201604020003, the Guangdong Province Key Laboratory of Computational Science Open Grant No. 2018009, the Construction Project of Shanghai Key Laboratory of Molecular Imaging 18DZ2260400, the Clinical Research Startup Program of Southern Medical University by High-level University Construction Funding of Guangdong Provincial Department of Education under Grant No. LC2016ZD018, and the Clinical Research Program of Nanfang Hospital, Southern Medical University under Grant No. 2018CR040, and President's fund of Nanfang Hospital, Southern Medical University under Grant No. 2019C017, in part by the Science and Technology Program of Guangzhou under Grant 201804020053.

Authors' contributions

ZH collected data, designed the experiments and wrote the manuscript in this study. Y.Li wrote the code, implement the experiments and wrote the manuscript. WX, CW, HZ, ZX, SW, J.Wu, CF and MM participated in observer study. XM and J.Wei provided the guidance for establishing deep learning model. WC, Y.Lu and GQ provided the guidance for the whole study including model design, experiment design and manuscript writing.

Acknowledgements

Zilong He and Yue Li contributed equally to the work. This work was supported, in part, by the National Key R&D Program of China under Grant Nos. 2018YFC1704206 and 2016YFB0200602, 2019YFC0121903 and 2019YFC0117301, the Fundamental Research Funds for the Central Universities under Grant No. 19LGYJS63, the NSFC under Grant Nos. 81971691, 81801809, 81830052, 81827802, U1811461, and 11401601, the Natural Science Foundation of Guangdong Province, China under Grant No. 2019A1515011168, and 2018A0303130215, the Science and Technology Innovative Project of Guangdong Province under Grant Nos. 2016B030307003, 2015B010110003, and 2015B020233008, the Science and Technology Planning Project of Guangdong Province under Key Grant Nos. 2015B020233002, and 2017B020210001, the Guangzhou Science and Technology Creative Project under Key Grant No. 201604020003, the Guangdong Province Key Laboratory of Computational Science Open Grant No. 2018009, the Construction Project of Shanghai Key Laboratory of Molecular Imaging 18DZ2260400, the Clinical Research Startup Program of Southern Medical University by High-level University Construction Funding of Guangdong Provincial

Department of Education under Grant No. LC2016ZD018, and the Clinical Research Program of Nanfang Hospital, Southern Medical University under Grant No. 2018CR040, and President's fund of Nanfang Hospital, Southern Medical University under Grant No. 2019C017, in part by the Science and Technology Program of Guangzhou under Grant 201804020053.

Footnotes

Not applicable

References

1. Smith RA, Cokkinides V, Brooks D, et al (2010) Cancer screening in the United States, 2010: a review of current American Cancer Society guidelines and issues in cancer screening. CA Cancer J Clin 60(2):99–119
2. Broeders M, Moss S, Nyström L et al (2012) The impact of mammographic screening on breast cancer mortality in Europe: a review of observational studies. J Med Screen 19:14–25
3. Independent UK Panel on Breast Cancer Screening (2012) The benefits and harms of breast cancer screening: an independent review. Lancet 380:1778–1786
4. Karssemeijer N, Bluekens AM, Beijerinck D et al (2009) Breast cancer screening results 5 years after introduction of digital mammography in a population-based screening program. Radiology 253:353–358
5. Lu S, Huang X, Yu H et al (2016) Dietary patterns and risk of breast cancer in Chinese women: a population-based case-control study. Lancet 388: S61
6. Karssemeijer N, Bluekens AM, Beijerinck D et al (2009) Breast cancer screening results 5 years after introduction of digital mammography in a population-based screening program. Radiology 253:353–358
7. Rimmer A (2017) Radiologist shortage leaves patient care at risk, warns Royal College. BMJ 359: j4683
8. EvansK K, Birdwell RL, Wolfe JM (2013) If you don't find it often, you often don't find it: why some cancers are missed in breast cancer screening. PLoS One 8: e64366
9. Bird RE, Wallace TW, Yankaskas BC (1992) Analysis of cancers missed at screening mammography. Radiology 184(3):613–617
10. Wong, C. et al (2011) Mammographic density and its interaction with other breast cancer risk factors in an Asian population. British journal of cancer 104, 871–874
11. El-Bastawissi, A. Y., White, E., Mandelson, M. T. et al (2001) Variation in mammographic breast density by race. Annals of epidemiology 11, 257–263.
12. Rajaram, N. et al (2017) Differences in mammographic density between Asian and Caucasian populations: a comparative analysis. Breast Cancer Research and Treatment 161, 353–362, https://doi.org/10.1007/s10549-016-4054-y
13 Friedewald, S. M. et al (2014) Breast cancer screening using tomosynthesis in combination with digital mammography. JAMA Jun 25;311(24):2499-507
14. Broeders MJ, Onland-Moret NC, Rijken HJ et al (2003) Use of previous screening mammograms to identify features indicating cases that would have a possible gain in prognosis following earlier detection. Eur J Cancer 39(12):1770–1775.
15. H. D. Cheng, X. J. Shi, R. Min, et al (2005) "Approaches for automated detection and classification of masses in mammograms," Pattern Recognition, vol. 39, no. 4, pp. 646–668, Apr. 2006, doi: 10.1016/j.patcog.07.006
16. B. Q. Huynh, H. Li, and M. L. Giger (2016) "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," J. Med. Imag, vol. 3, no. 3, p. 034501, Aug
17. A. Rodríguez-Ruiz et al (2019) "Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System," Radiology, vol. 290, no. 2, pp. 305–314

18. N. Wu et al (2020) “Deep Neural Networks Improve Radiologists’ Performance in Breast Cancer Screening,” IEEE Trans. Med. Imaging, vol. 39, no. 4, pp. 1184–1194
19. Z. Jiao, X. Gao, Y. Wang et al (2016) “A deep feature based framework for breast masses classification,” Neurocomputing, vol. 197, pp. 221–231
20. T. Kooi et al (2016) “Large scale deep learning for computer aided detection of mammographic lesions,” Medical Image Analysis, vol. 35, pp. 303–312
21. Karen Simonyan, Andrew Zisserman (2015) Very Deep Convolutional Networks for Large-scale Image Recognition, http://arXiv.org/cs/ 1409.1556. Accessed 10 Apr 2015
22. Alejandro Rodríguez-Ruiz, MSc • Elizabeth Krupinski, PhD • Jan-Jurre Mordang, MSc et al (2019) Detection of Breast Cancer with Mammography:Effect of an Artificial Intelligence Support System, Radiology 00:1–10
23. Lehman CD, Wellman RD, Buist DS et al (2015) Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med 175:1828–1837
24. Krzysztof J Geras, Ritse M Mann, Linda Moy et al (2019) Artificial Intelligence for Mammography and Digital Breast Tomosynthesis: Current Concepts and Future Perspectives, Radiology 293(2):246-259
25. Kim EK, Kim HE, Han K, et al (2018) Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study. Sci Rep 8(1):2762
26. Alejandro Rodriguez-Ruiz, Kristina Lång, Albert Gubern-Merida et al (2019) Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study, European Radiology 29:4825–4832

Figures

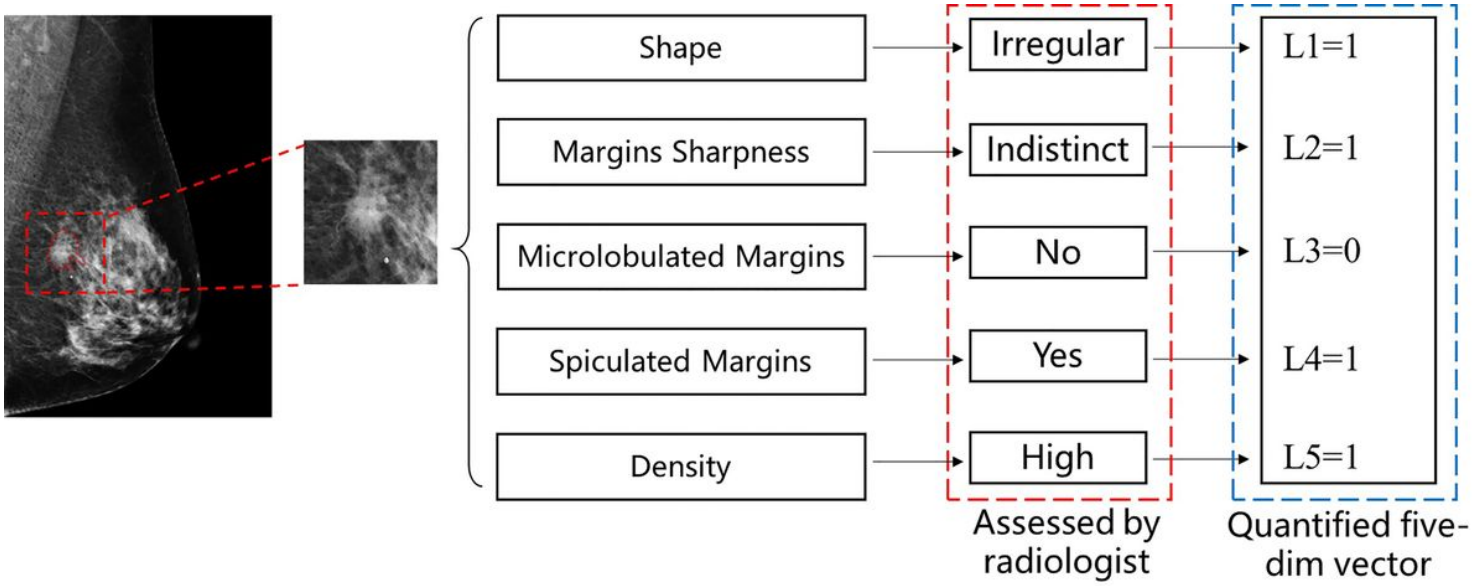

Figure 1

An example of quantification for a malignant mass in MLO-view FFDM. Five text descriptions assessed by radiologist as shown in red box are quantified as corresponding numbers. A five-dimension vector is generated, which is used as the ground-truth to train the perceptive feature extractor.

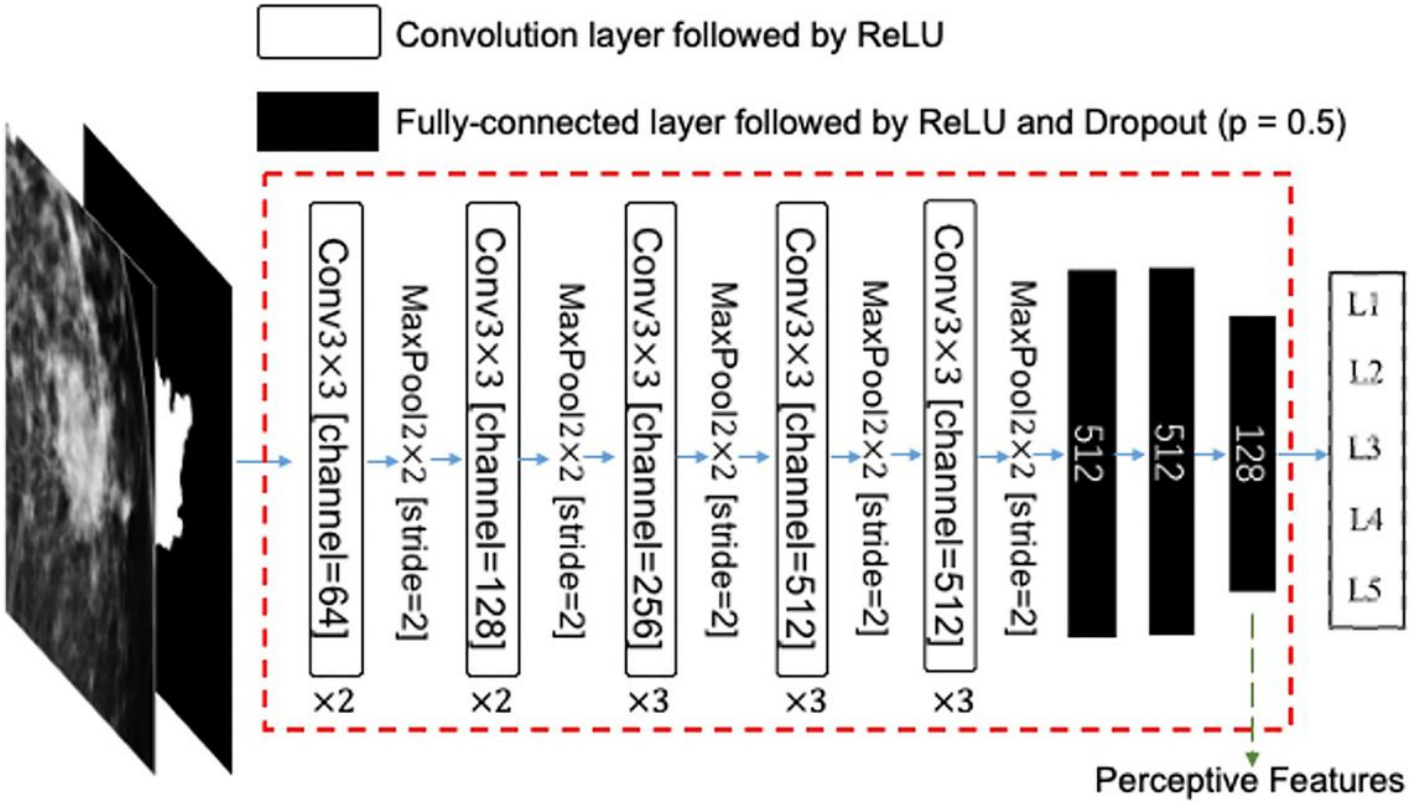

Figure 2

The architecture of this feature extractor.

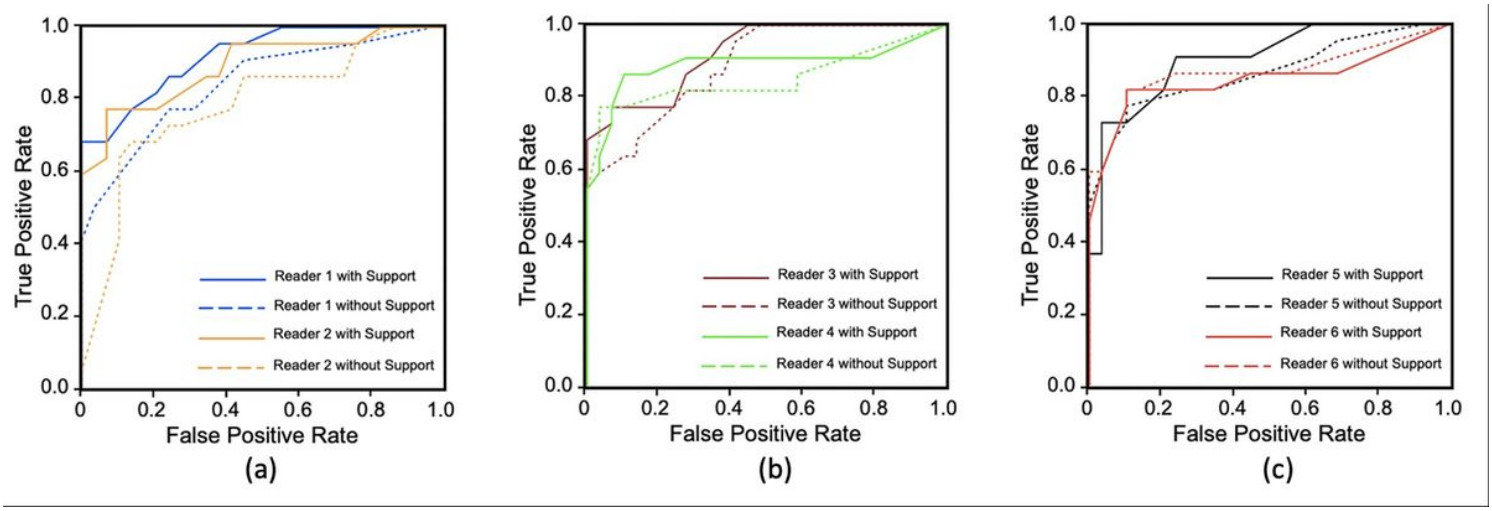

Figure 3

ROC curves for six readers in three groups diagnosis with and without CAD model support. (a)Junior group of reader 1 and 2; (b) Middle-seniority group of reader 3 and 4; (c) Senior group of reader 5 and 6.

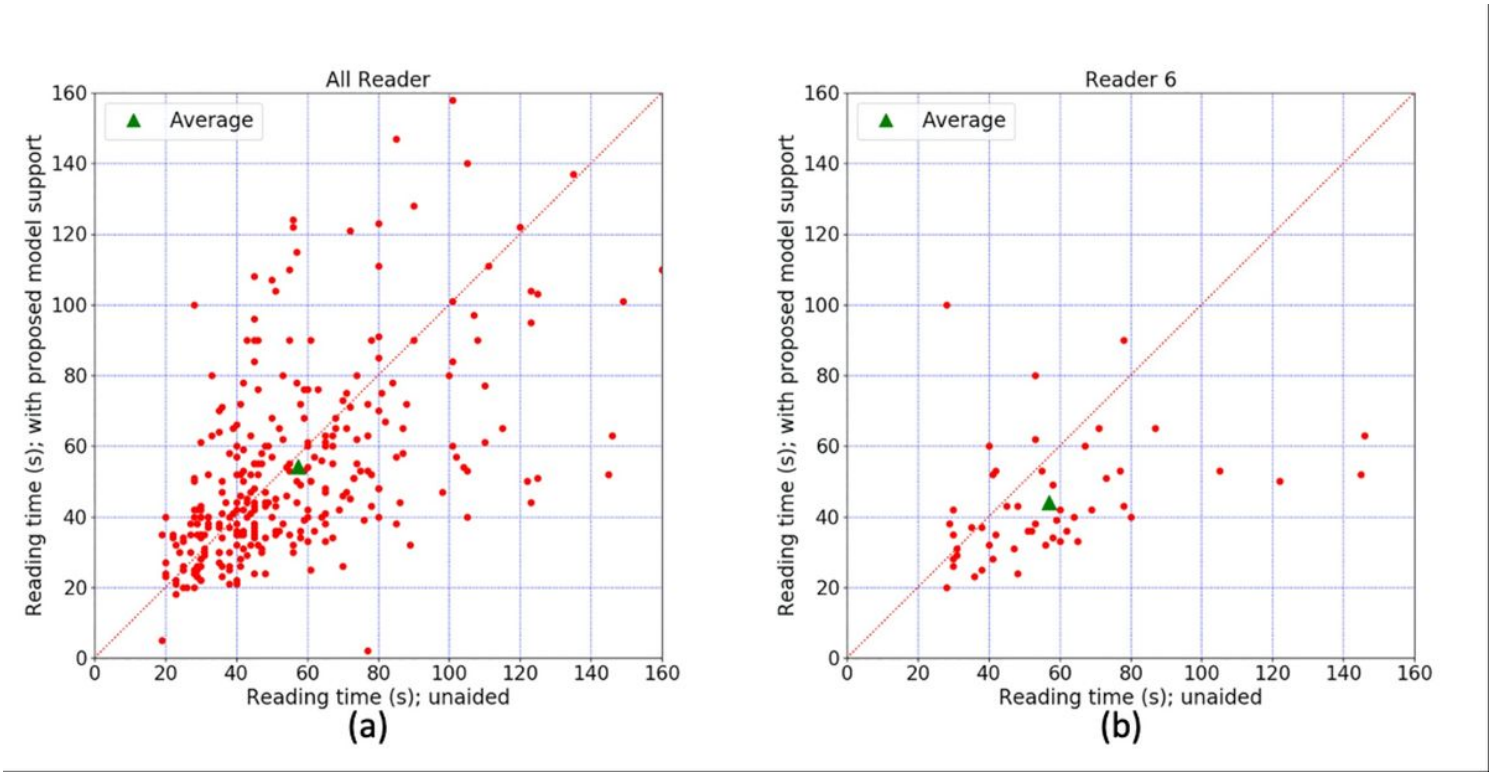

Figure 4

Diagnosis time comparison. (a) The time comparison of all readers; (b) the time comparison for reader 6, who was the only one that showed obvious difference between two sessions. Graph shows differences in diagnosis time per case for all reader. Each red point indicated diagnosis time for a certain case with or without model support. There is no significant change when the point falls on diagonal. Point above the

diagonal indicates diagnosis time has increase with model support. Point below the diagonal means the time decrease with model support.

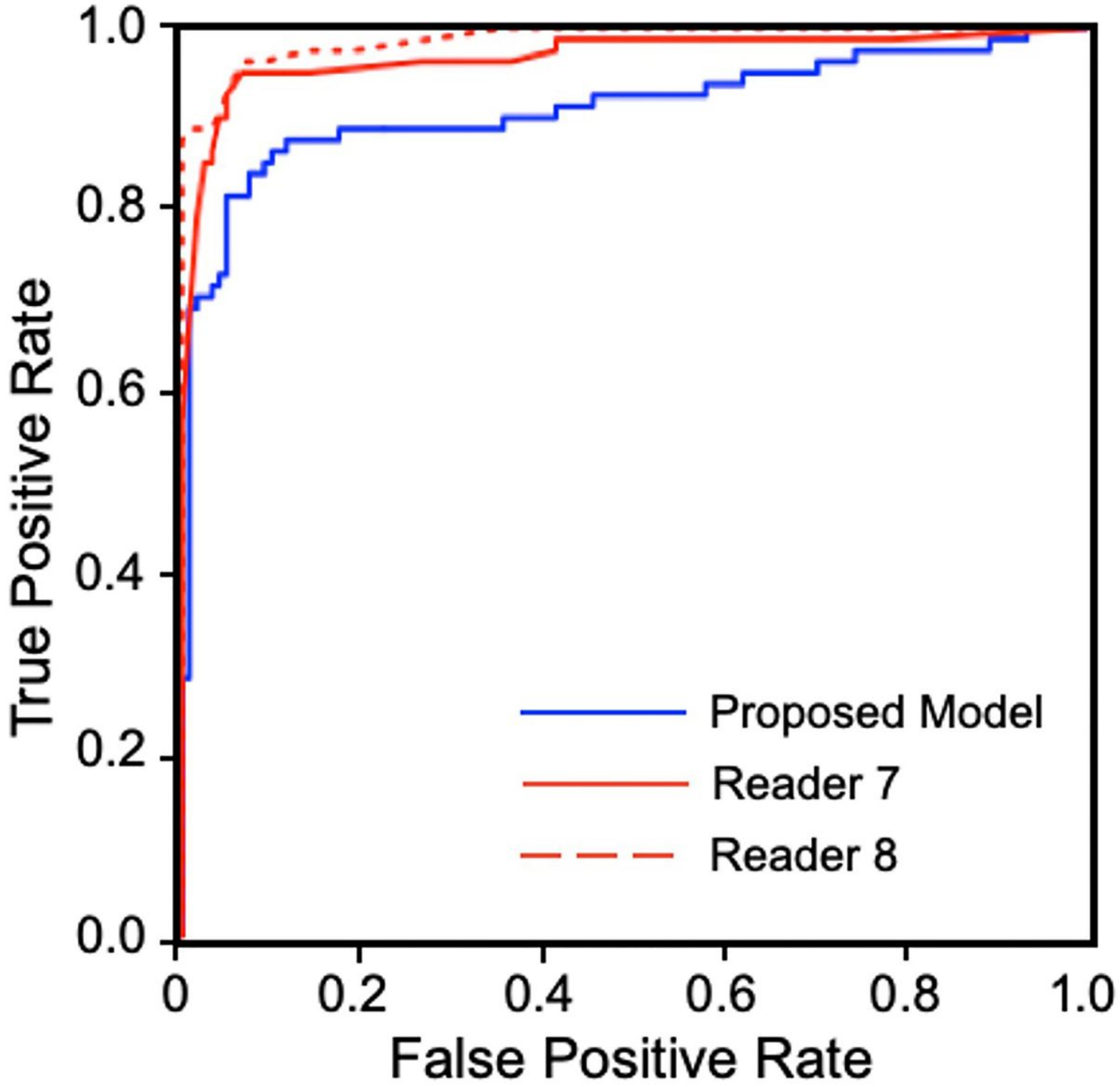

Figure 5

The ROC curves of radiologists and our proposed model in 202 independent test cases.