Thinking Like Sonographers: A Deep CNN Model for Diagnosing Gout from Musculoskeletal Ultrasound

Zhi Cao[1], Weijing Zhang[2], Keke Chen[2], Di Zhao[2], Daoqiang Zhang[1], Hongen Liao[3], and Fang Chen[1(✉)]

[1] Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
chenfang@nuaa.edu.cn

[2] Department of Ultrasound, Nanjing Drum Tower Hosptial, The Affiliated Hospital of Nanjing University Medical School, Nanjing 210008, China

[3] Department of Biomedical Engineering, School of Medicine, Tsinghua University, Beijing 100084, China

Abstract. We explore the potential of deep convolutional neural network (CNN) models for differential diagnosis of gout from musculoskeletal ultrasound (MSKUS), as no prior study on this topic is known. Our exhaustive study of state-of-the-art (SOTA) CNN image classification models for this problem reveals that they often fail to learn the gouty MSKUS features, including the double contour sign, tophus, and snowstorm, which are essential for sonographers' decisions. To address this issue, we establish a framework to adjust CNNs to "think like sonographers" for gout diagnosis, which consists of three novel components: (1) Where to adjust: Modeling sonographers' gaze map to emphasize the region that needs adjust; (2) What to adjust: Classifying instances to systematically detect predictions made based on unreasonable/biased reasoning and adjust; (3) How to adjust: Developing a training mechanism to balance gout prediction accuracy and attention reasonability for improved CNNs. The experimental results on clinical MSKUS datasets demonstrate the superiority of our method over several SOTA CNNs.

Keywords: Musculoskeletal ultrasound · Gout diagnosis · Gaze tracking · Reasonability

1 Introduction

Gout is the most common inflammatory arthritis and musculoskeletal ultrasound (MSKUS) scanning is recommended to diagnose gout due to the non-ionizing radiation, fast imaging speed, and non-invasive characteristics of MSKUS [7]. However, misdiagnosis of gout can occur frequently when a patient's clinical characteristics are atypical. Traditional MSKUS diagnosis relies on the experience of the radiologist which is time-consuming and labor-intensive. Although

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43987-2_16.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 159–168, 2023.
https://doi.org/10.1007/978-3-031-43987-2_16

convolutional neural networks (CNNs) based ultrasound classification models have been successfully used for diseases such as thyroid nodules and breast cancer, conspicuously absent from these successful applications is the use of CNNs for gout diagnosis from MSKUS images.

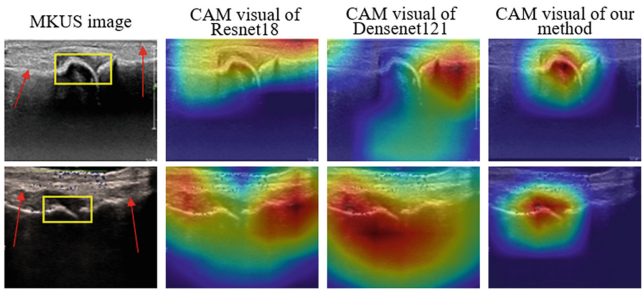

Fig. 1. (a) MSKUS images. Yellow boxes denote the gaze areas of the sonographers and red arrows denote the surrounding fascial tissues; (b) Grad-Cam visual of ResNet18; (c) Grad-Cam visual of Densenet121; (d) Grad-Cam visual of our method. (Color figure online)

There are significant challenges in CNN based gout diagnosis. Firstly, the gout-characteristics contain various types including double contour sign, synovial hypertrophy, synovial effusion, synovial dislocation and bone erosion, and these gout-characteristics are small and difficult to localize in MSKUS. Secondly, the surrounding fascial tissues such as the muscle, sarcolemma and articular capsule have similar visual traits with gout-characteristics, and we found the existing CNN models can't accurately pay attention to the gout-characteristics that radiologist doctors pay attention to during the diagnosis process (as shown in Fig. 1). Due to these issues, SOTA CNN models often fail to learn the gouty MSKUS features which are key factors for sonographers' decision.

In medical image analysis, recent works have attempted to inject the recorded gaze information of clinicians into deep CNN models for helping the models to predict correctly based on lesion area. Mall et al. [9,10] modeled the visual search behavior of radiologists for breast cancer using CNN and injected human visual attention into CNN to detect missing cancer in mammography. Wang et al. [15] demonstrated that the eye movement of radiologists can be a new supervision form to train the CNN model. Cai et al. [3,4] developed the SonoNet [1] model, which integrates eye-gaze data of sonographers and used Generative Adversarial Networks to address the lack of eye-gaze data. Patra et al. [11] proposed the use of a teacher-student knowledge transfer framework for US image analysis, which combines doctor's eye-gaze data with US images as input to a large teacher model, whose outputs and intermediate feature maps are used to condition a student model. Although these methods have led to promising results, they can be difficult to implement due to the need to collect doctors' eye movement data for each image, along with certain restrictions on the network structure.

Different from the existing studies, we propose a novel framework to adjust the general CNNs to "think like sonographers" from three different levels. (1) Where to adjust: Modeling sonographers' gaze map to emphasize the region that needs adjust; (2) What to adjust: Classify the instances to systemically detect predictions made based on unreasonable/biased reasoning and adjust; (3) How to adjust: Developing a training mechanism to strike the balance between gout prediction accuracy and attention reasonability.

2 Method

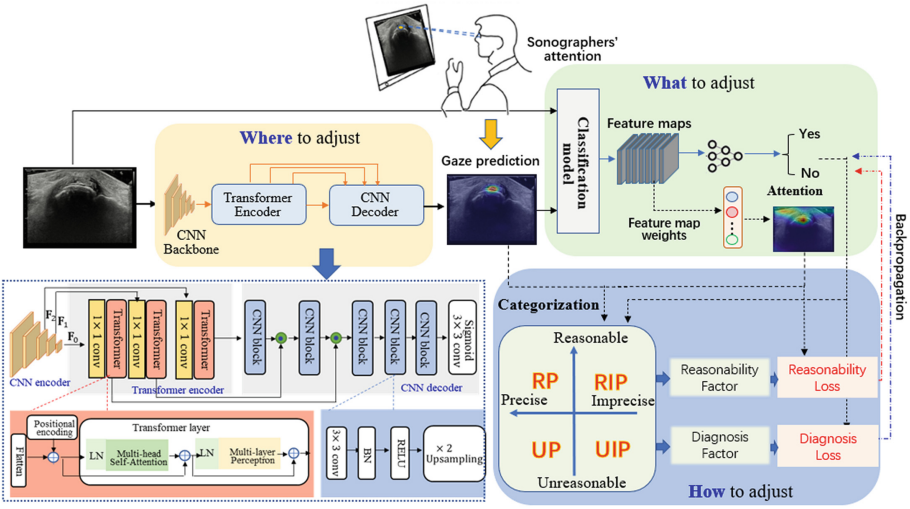

Fig. 2. The overall framework of the proposed method.

Figure 2 presents the overall framework, which controls CNNs to "think like sonographers" for gout diagnosis from three levels. 1) Where to adjust: we model the sonographers' gaze map to emphasize the region that needs control. This part learns the eye gaze information of the sonographers which is collected by the Eye-Tracker. 2) What to adjust: we divide instances into four categories to reflect whether the model prediction given to the instance is reasonable and precise. 3) How to adjust: a training mechanism is developed to strike the balance between gout diagnosis and attention accuracy for improving CNN.

2.1 Where to Adjust

It is essential to obtain the gaze map corresponding to each MSKUS to emphasize the region where gouty features are obvious. Inspired by studies of saliency

model [8], we integrate transformer into CNNs to capture multi-scale and long-range contextual visual information for modeling sonographers' gaze map. This gaze map learns the eye gaze information, collected by the Eye-Tracker, of the sonographers when they perform diagnosis. As shown in Fig. 2, this part consists of a CNN encoder for extracting multi-scale feature, a transformer-encoder for capturing long-range dependency, and a CNN decoder for predicting gaze map.

The MSKUS image $I_0 \in \mathbf{R}^{H \times W \times 3}$ is first input into CNN encoder that contains five convolution blocks. The output feature maps from the deeper last three convolution blocks are denoted as $\mathbf{F}_0, \mathbf{F}_1, \mathbf{F}_2$ and are respectively fed into transformer encoders to enhance the long-range and contextual information. During the transformer-encoder, we first flatten the feature maps produced by the CNN encoder into a 1D sequence. Considering that flatten operation leads to losing the spatial information, the absolute position encoding [14] is combined with the flatten feature map via element-wise addition to form the input of the transformer layer. The transformer layer contains the standard Multi-head Self-Attention (MSA) and Multi-layer Perceptron (MLP) blocks. Layer Normalization (LN) and residual connection are applied before and after each block respectively.

In the CNN decoder part, a pure CNN architecture progressively up-samples the feature maps into the original image resolution and implements pixel-wise prediction for modeling sonographers' gaze map. The CNN decoder part includes five convolution blocks. In each block, 3×3 convolution operation, Batch normalization (BN), RELU activation function, and 2-scale upsampling that adopts nearest-neighbor interpolation is performed. In addition, the transformer's output is fused with the feature map from the decoding process by an element-wise product operation to further enhance the long-range and multi-scale visual information. After five CNN blocks, a 3×3 convolution operation and Sigmoid activation is performed to output the predicted sonographers' gaze map. We use the eye gaze information of the sonographers which is collected by the Eye-Tracker to restrain the predicted sonographers' gaze map. The loss function is the sum of the Normalized Scanpath Saliency (NSS), the Linear Correlation Coefficient (CC), Kullback-Leibler divergence (KLD) and Similarity (SIM) [2].

2.2 What to Adjust

Common CNN classification models for gout diagnosis often fail to learn the gouty MSKUS features including the double contour sign, tophus, and snowstorm which are key factors for sonographers' decision. A CAM for a particular category indicates the discriminative regions used by the CNN to identify that category. Inspired by CAM technique, it is needed to decide whether the attention region given to an CNN model is reasonable for diagnosis of gout. We firstly use the Grad-CAM technique [12] to acquire the salient attention region S_{CAM} that CNN model perceives for differential diagnosis of gout. To ensure the scale of the attention region S_{CAM} is the same as the sonographers' gaze map S_{sono} which is modeled by saliency model, we normalize S_{CAM} to the values between 0 and 1, get $\widetilde{S}_{CAM}$. Then we make bit-wise intersection over union(IoU) operations with the S_{sono} and $\widetilde{S}_{CAM}$ to measure how well

the two maps overlap. Note that we only calculate the part of $\widetilde{S}_{CAM}$ that is greater than 0.5. For instances whose IoU is less than 50%, we consider that the model's prediction for that instance is unreasonable. As shown in Fig. 3, when CNN do prediction, we can divide the instances into four categories:

RP: Reasonable Precise: The attention region focusses on the gouty features which are important for sonographers' decision, and the diagnosis is precise.
RIP: Reasonable Imprecise: Although attention region focusses on the gouty features, while the diagnosis result is imprecise.
UP: Unreasonable Precise: Although the gout diagnosis is precise, amount of attention is given to irrelevant feature of MSKUS image.
UIP: Unreasonable Imprecise: The attention region focusses on irrelevant features, and the diagnosis is imprecise.

Our target of adjustment is to reduce imprecise and unreasonable predictions. In this way, CNNs not only finish correct gout diagnosis, but also acquire the attention region that agreements with the sonographers' gaze map.

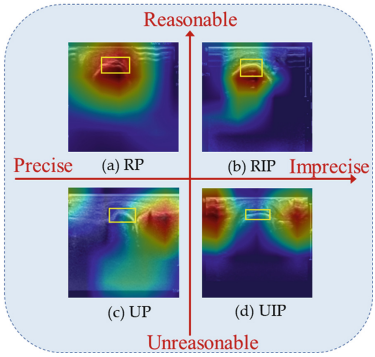

Fig. 3. Four categories: (a) RP (b) RIP (c) UP (d) UIP. Yellow boxes denote the gaze areas of the sonographers.

2.3 How to Adjust

We proposed a training mechanism (Algorithm 1) which can strike the balance between the gout diagnosis error and the reasonability error of attention region to promote the CNNs to "think like sonographers". In addition to reducing the diagnosis error, we also want to minimize the difference between sonographers' gaze map S_{sono} and normalized salient attention region $\widetilde{S}_{CAM}$, which directly leads to our target:

Algorithm 1: Proposed training mechanism

Input:D_{train}, O_{target}, M_{attn} generated from saliency model,
W_{base} parameters of base model needed to be adjusted
Output:Optimized model parameters W^*
$O_{pred}, M_{cam} = model(D_{train}|W_{base})$;
split D_{train} into 4 categories based on (O_{pred}, O_{target}) and (M_{cam}, M_{attn});
set the α base on the ratio of 4 categories;
$W^* = W_{base}$;
for $epoch = \{1 \ldots, N\}$ **do**
    $W^* = train(W^*, D_{train}, M_{attn}, \alpha)$;
end

$$\boldsymbol{min}(\boldsymbol{L}_{diagnosis} + \boldsymbol{L}_{reasonability})$$

$$\boldsymbol{L}_{reasonability} = L_1(\widetilde{S}_{CAM}, S_{sono})$$

The total loss function can be expressed as the weighted sum the gout diagnosis error and the reasonability error, as follows:

$$\boldsymbol{L}_{total} = \alpha \boldsymbol{L}_{diagnosis} + (1 - \alpha)\boldsymbol{L}_{reasonability}$$

The gout diagnosis error $\boldsymbol{L}_{diagnosis}$ is calculated by the Cross-entropy loss, and the reasonability is calculated by the L1-loss. This training mechanism uses the quadrant of instances to identify whether samples' attention needs to adjusted. For MSKUS sample in the quadrant of UP, α can be set 0.2 to control the CNN pay more attention to reasonability. Correspondingly, for sample in RIP, α can be set 0.8 to make CNN pay more attention to precise. For sample in RP and UIP, α can be set 0.5 to strike the balance between accuracy and reasonability.

3 Experiments

MSKUS Dataset Collection. The MSKUS data were collected for patients suspected of metatarsal gout in Nanjing Drum Tower Hosptial. Informed written consent was obtained at the time of recruitment. Dataset totally contains 1127 US images from different patients including 509 gout images and 618 healthy images. The resolution of the MSKUS images were resized to 224 × 224. During experiments, we randomly divided 10% of the dataset into testing sets, then the remaining data was divided equally into two parts for the different phases of the training. We used 5-fold cross validation to divide the training sets and validation sets.

Gaze Data Collection. We collected the eye movement data with the Tobii 4C eye-tracker operating at 90 Hz. The MSKUS images were displayed on a 1920 × 1080 27-inch LCD screen. The eye tracker was attached beneath the screen with a magnetic mounting bracket. Sonographers were seated in front of the screen and free to adjust the chair's height and the display's inclination. Binary maps of the same size as the corresponding MSKUS images were generated using the gaze data, with the pixel corresponding to the point of gaze marked with a'1' and the other pixels marked with a'0'. A sonographer gaze map S was generated for each binary map by convolving it with a truncated Gaussian Kernel G($\sigma_{x,y}$), where G has 299 pixels along x dimension, and 119 pixels along y dimension.

Evaluation Metrics. Five metrics were used to evaluate model performance: Accuracy (ACC), Area Under Curve (AUC), Correlation Coefficient (CC), Similarity (SIM) and Kullback-Leibler divergence (KLD) [2]. ACC and AUC were implemented to assess the gout classification performance of each model, while CC, SIM, and KLD were used to evaluate the similarity of the areas that the model and sonographers focus on during diagnoses.

Evaluation of "Thinking like Sonographers" Mechanism. To evaluate the effectiveness of our proposed mechanism of "Thinking like Sonographers" (TLS) that combines "where to adjust", "what to adjust" and "how to adjust", we compared the gout diagnosis results of several classic CNN classification [5, 6, 13] models without/with our TLS mechanism. The results, shown in Table 1, revealed that using our TLS mechanism led to a significant improvement in all metrics. Specifically, for ACC and AUC, the model with our TLS mechanism achieved better results than the model without it. Resnet34 with TLS acquired the highest improvement in ACC with a 4.41% increase, and Resnet18 with TLS had a 0.027 boost in AUC. Our TLS mechanism consistently performed well in improving the gout classification performance of the CNN models. More comparison results were shown in Appendix Fig. A1 and Fig. A2.

Table 1. The performances of models training wi/wo our mechanism in MSKUS.

Method	ACC↑ (%)	AUC↑	CC↑	SIM↑	KLD↓
Resnet18 wo TLS	86.46 ± 3.90	0.941 ± 0.023	0.161 ± 0.024	0.145 ± 0.011	3.856 ± 0.235
Resnet18 wi TLS	89.13 ± 3.32	$\mathbf{0.968 \pm 0.005}$	0.404 ± 0.004	0.281 ± 0.003	$\mathbf{1.787 \pm 0.017}$
Resnet34 wo TLS	83.15 ± 3.78	0.922 ± 0.024	0.190 ± 0.054	0.151 ± 0.020	3.500 ± 0.530
Resnet34 wi TLS	87.56 ± 1.89	0.947 ± 0.018	0.376 ± 0.006	0.252 ± 0.004	1.951 ± 0.043
Resnet50 wo TLS	88.82 ± 1.16	0.956 ± 0.008	0.189 ± 0.024	0.157 ± 0.013	4.019 ± 0.522
Resnet50 wi TLS	89.61 ± 3.05	0.967 ± 0.011	0.402 ± 0.028	0.298 ± 0.020	2.133 ± 0.232
Vgg16 wo TLS	89.13 ± 3.20	0.958 ± 0.021	0.221 ± 0.089	0.182 ± 0.044	3.461 ± 0.776
Vgg16 wi TLS	$\mathbf{91.50 \pm 2.88}$	0.966 ± 0.020	$\mathbf{0.416 \pm 0.020}$	$\mathbf{0.305 \pm 0.013}$	1.932 ± 0.084
DenseNet121 wo TLS	88.82 ± 2.08	0.956 ± 0.015	0.175 ± 0.030	0.152 ± 0.011	3.822 ± 0.599
DenseNet121 wi TLS	89.45 ± 1.62	0.965 ± 0.010	0.368 ± 0.011	0.239 ± 0.007	1.991 ± 0.062

The CC, SIM, and KLD metrics were utilized to assess the similarity between the CAMs of classification models and the collected gaze maps, providing an indication of whether the model was able to "think" like a sonographer. Table 1 showed that the models with our TLS mechanism achieved significantly better results in terms of CC and SIM (i.e., higher is better), as well as a decline of more than 1.50 in KLD (lower is better), when compared to the original models. This indicated that the models with TLS focused on the areas shown to be similar to the actual sonographers. Furthermore, Fig. 4 illustrated the qualitative results of CAMs of models with and without TLS mechanism. The original models without TLS paid more attention to noise, textures, and artifacts, resulting in unreasonable gout diagnosis. With TLS, however, models could focus on the crucial areas in lesions, allowing them to think like sonographers.

Stability Under Different Gaze Maps via t-Test. To evaluate the prediction's stability under the predicted gaze map from the generation model in "Where to adjust", we conducted three t-test studies. Specifically, we trained two classification models (M_C and M_P), using the actual collected gaze maps, and the predicted maps from the generation model, respectively. During the testing, we used the collected maps as input for M_C and M_P to get classification results R_{CC} and R_{PC}. Similarly, we used the predicted maps as input for M_C andM_P

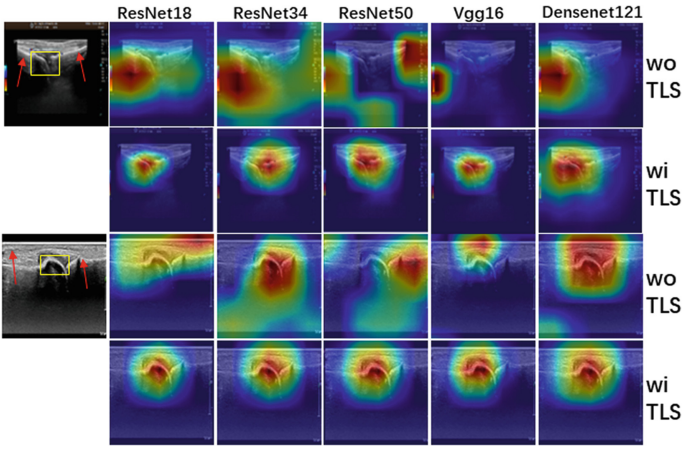

Fig. 4. Grad-CAM for ResNet18, ResNet34, ResNet50, vgg16 and Densenet121. Yellow boxes denote sonographers' gaze areas and red arrows denote the fascial tissues. (Color figure online)

to get the results R_{CP} and R_{PP}. Then, we conducted three T-tests: (1) between R_{CC} and R_{PC}; (2) between R_{CP} and R_{PP}; and (3) between R_{CC} and R_{PP}.

As shown in Table 2, the p-values of t-test (1)(2) and (3) are all greater than 0.005, suggesting that no significant difference was observed between the classification results obtained from different generative strategies. This implied that our training mechanism was model-insensitive. Consequently, it was possible to use predicted gaze maps for both the training and testing phases of the classification models without any notable performance decrease. This removed the need to collect eye movement maps during the training and testing phases, significantly lightening the workload of data collection. Therefore, our TLS mechanism, which involved predicting the gaze maps, could potentially be used in clinical environments. This would allow us to bypass the need to collect the real gaze maps of the doctors while classifying newly acquired US images, and thus improved the clinical implications of our mechanism, "Thinking like Sonographers".

Table 2. Statistical test results

Method/p-value	ResNet18	ResNet34	ResNet50	Vgg16	DenseNet121
t-test(1)	0.4219	0.8719	0.8701	0.6281	0.4428
t-test(2)	0.4223	0.8700	0.8725	0.6272	0.4434
t-test(3)	0.4192	0.8714	0.8727	0.6191	0.4446

4 Conclusion

In this study, we propose a framework to adjust CNNs to "think like sonographers", and diagnose gout from MSKUS images. The mechanism of "thinking like sonographers" contains three levels: where to adjust, what to adjust, and how to adjust. The proposed design not only steers CNN models as we intended, but also helps the CNN classifier focus on the crucial gout features. Extensive experiments show that our framework, combined with the mechanism of "thinking like sonographers" improves performance over the baseline deep classification architectures. Additionally, we can bypass the need to collect the real gaze maps of the doctors during the classification of newly acquired MSKUS images, thus our method has good clinical application values.

Acknowledgment. This work was supported in part by National Nature Science Foundation of China grants (62271246, U20A20389, 82027807, U22A2051), Key Research and Development Plan of Jiangsu Province (No. BE2022842), National Key Research and Development Program of China (2022YFC2405200).

References

1. Baumgartner, C.F., et al.: Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Trans. Med. Imaging **36**(11), 2204–2215 (2017)
2. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE Trans. Pattern Anal. Mach. Intell. **41**(3), 740–757 (2018)
3. Cai, Y., Sharma, H., Chatelain, P., Noble, J.A.: Multi-task SonoEyeNet: detection of fetal standardized planes assisted by generated sonographer attention maps. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 871–879. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_98
4. Cai, Y., Sharma, H., Chatelain, P., Noble, J.A.: Sonoeyenet: standardized fetal ultrasound plane detection informed by eye tracking. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1475–1478. IEEE (2018)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
6. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
7. Liu, S., et al.: Deep learning in medical ultrasound analysis: a review. Engineering **5**(2), 261–275 (2019)
8. Lou, J., Lin, H., Marshall, D., Saupe, D., Liu, H.: Transalnet: towards perceptually relevant visual saliency prediction. Neurocomputing **494**, 455–467 (2022)
9. Mall, S., Brennan, P.C., Mello-Thoms, C.: Modeling visual search behavior of breast radiologists using a deep convolution neural network. J. Med. Imaging **5**(3), 035502–035502 (2018)

10. Mall, S., Krupinski, E., Mello-Thoms, C.: Missed cancer and visual search of mammograms: what feature-based machine-learning can tell us that deep-convolution learning cannot. In: Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment, vol. 10952, pp. 281–287. SPIE (2019)
11. Patra, A., et al.: Efficient ultrasound image analysis models with sonographer gaze assisted distillation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 394–402. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_43
12. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
14. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
15. Wang, S., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Follow my eye: using gaze to supervise computer-aided diagnosis. IEEE Trans. Med. Imaging **41**(7), 1688–1698 (2022)