Follow Sonographers' Visual Scan-path: Adjusting CNN Model for Diagnosing Gout from Musculoskeletal Ultrasound

Xin Tang[1] ⋆, Zhi Cao[1] ⋆, Weijing Zhang[2], Di Zhao[2], Hongen Liao[3,4], Daoqiang Zhang[1], Fang Chen[3](✉)

[1] College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing 211106, China

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

[3] School of Biomedical Engineering and the Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, 200240, China
chenfang_bme@163.com

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

Abstract. The current models for automatic gout diagnosis train convolutional neural network (CNN) using musculoskeletal ultrasound (MSKUS) images paired with classification labels, which are annotated by experience sonographers. However, this prevalent diagnostic model overlooks valuable supplementary information derived from sonographers' annotations, such as the visual scan-path followed by sonographers. We notice that this annotation procedure offers valuable insight into human attention, aiding the CNN model in focusing on crucial features in gouty MSKUS scans, including the double contour sign, tophus, and snowstorm, which play a crucial role in sonographers' diagnostic decisions. To verify this, we create a gout MSKUS dataset that enriched with sonographers' annotation byproduct visual scan-path. Furthermore, we introduce a scan-path based fine-tuning training mechanism (SFT) for gout diagnosis models, leveraging the annotation byproduct scan-paths for enhanced learning. The experimental results demonstrate the superiority of our SFT method over several SOTA CNNs.

Keywords: Musculoskeletal ultrasound · Gout diagnosis · Visual scan-path.

1 Introduction

Gout, a typical manifestation of inflammatory arthritis, is diagnosed through the identification of monosodium urate crystals in aspirate from synovial fluid

⋆ Xin Tang and Zhi Cao contribute equally to this work.

or tophi. However, the use of arthrocentesis for this purpose is invasive and not always practical for all patients [10]. Musculoskeletal ultrasound (MSKUS), known for its real-time, non-invasive, and high-resolution capabilities, has emerged as a potent tool for evaluating the joint status of individuals with gout [13]. Nevertheless, the accuracy of MSKUS-based gout diagnosis heavily depends on the experience of the sonographers, introducing subjectivity and the potential for misdiagnosis when the patient's clinical manifestations are challenging to differentiate. The diagnosis process is subjective and time-consuming, so it is necessary to introduce automatic diagnosis system. While automatic ultrasound diagnosis models based on deep CNNs [7,14] have been extensively explored for diseases like thyroid nodules and breast cancer, significant challenges persist in the realm of gout diagnosis. Existing studies indicate that common gout features on MSKUS include synovial effusion, the double contour sign, and gout stones. Using synovial effusion as an example, it is characterized by non-echoic or hypoechoic joint cavity widening in the MSKUS image [10,13]. The feature region is small, contains limited information, and bears similarity to other regional features (as illustrated in Fig. 1 a). This similarity poses a challenge for current CNN models to effectively extract the features with the Class Activation Map (CAM) [9] method in the MSKUS image (as depicted in Fig. 1 c).

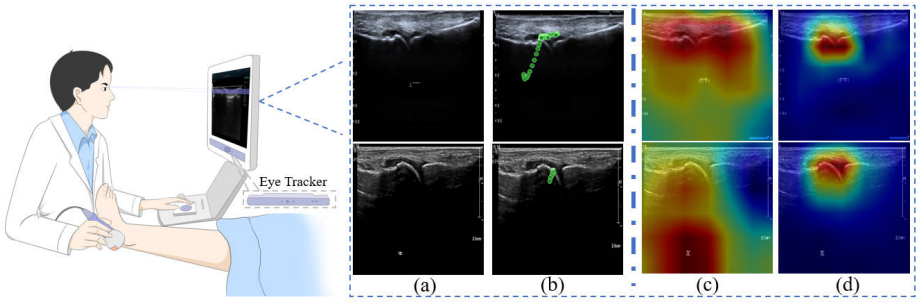

Fig. 1. (a) MSKUS images; (b) the sonographers' scan path in annotation procedure; (c) Grad-CAM of current CNN model (Using DenseNet121 as an example); (d) Grad-CAM of our model.

To address these challenges, recent efforts have explored the use of human visual attention for lesion diagnosis in ultrasound images. For example, Alsharid et al. [1] introduced multi-modal deep neural networks for analyzing fetal ultrasound videos, incorporating sonographer gaze in the form of attention maps. Cai et al. [4] proposed the SonoNet [2] model, which integrates the attention map of sonographers into different convolution layers to enhance recognition accuracy. While these methods have shown promise, they require collecting eye movement data for each tested ultrasound image in real clinic applications. Cao et al. [6] presented a framework for adjusting CNNs to "think like sonographers" in gout diagnosis, involving a comparison between sonographers' attention maps and

those generated by CNNs. This model strikes a balance between accuracy and reasonability during the training phase. However, existing studies used static methods to model the gaze point of each image to generate corresponding visual heatmaps, ignoring the learning of sonographers' eyes scanning pattern.

We noticed that the sonographers' annotation procedure mirrors the diagnostic process. In the procedure, the sonographer subconsciously searches the lesion features in the image. The sonographers' eye scan-path, mouse track, and other annotation byproducts generated during the diagnostic process serve as specific expressions of their professional experience (as depicted in Fig. 1b). Study has indicated that eye movement trajectories or eye scan-paths are closely linked to physicians' diagnostic outcomes and contain vital information for the diagnostic classification of images [11]. Therefore, the eye scan-path is more representative in annotation byproducts. Different from existing studies, we propose a novel framework——scan-path based fine-tuning training mechanism (SFT), to adjust the model to learn the sonographers' clinical experience for gout diagnosis.

Our contributions are outlined as follows:

1) We design a novel learnable kernel to recognize different sonographers' eyes scanning pattern.
2) Our proposed training mechanism (SFT) enables the CNN model to simultaneously learn gout diagnosis and capture the sonographer's attention.
3) Our training mechanism based on scan-paths demonstrates strong generalization and can be seamlessly integrated into various CNN backbones.

2 Materials and Methods

2.1 Sonographers' Visual Scan-path

Collecting Sonographers' Visual Scan-path. Eye tracking is achieved using a remote eye-tracker (Tobii Eyetracking Eye Track-er 4C, Danderyd, Sweden) mounted below the ultrasound machine display monitor. For each MSKUS frame, we record the visual scan-path of sonographer gaze. The eye tracker was placed on the lower side of the monitor. Sonographers do not have any visual or other signal to know that the eye-tracking device is functioning. Before the experiment begins, Sonographers perform a five-point calibration to ensure accurate eye movement data collection.

Processing Sonographers' Visual Scan-path. The visual scan-path data $\boldsymbol{G}$ recorded by an eye tracker for each MSKUS image frame can be represented as a sequence of tuples, denoted as $\boldsymbol{G} = \{g_i \in \mathbb{R}^K\}_{i=1}^N$. Each $g_i = (p_i^x, p_i^y, t_i, v_i)$ consists of the plane position coordinates (p_i^x, p_i^y) of the i-th gaze point, the corresponding time stamp t_i, and the validity v_i of whether the data is an outlier. Although the eye tracker can identify valid data samples, the raw data still contains noise and bias. Therefore, preprocessing of the raw scan-path data is

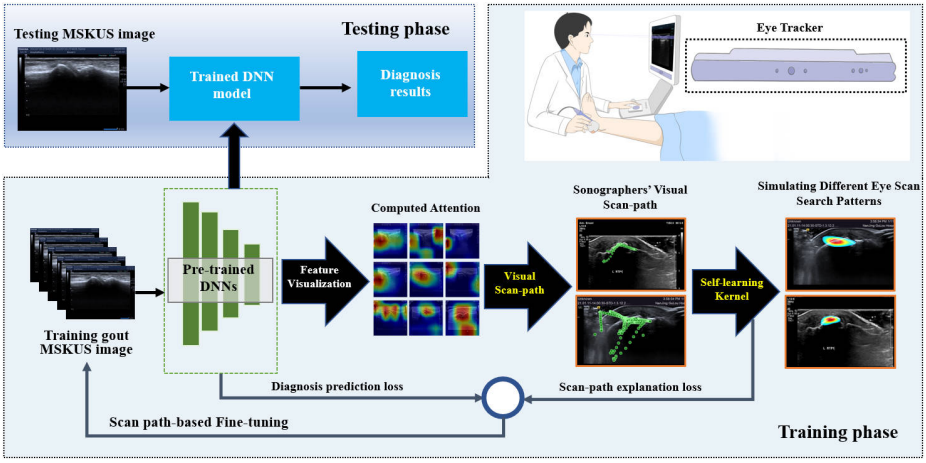

Fig. 2. The architecture of the scan-path based fine-tuning training mechanism.

necessary to obtain the fixation data of Sonographers. Firstly, the first 200ms of each sequence are disregarded to eliminate any bias caused by residual visual positions. Subsequently, a moving average filter with a window size of 3 samples is applied to mitigate high-frequency noise from eye tremors. To identify fixations, any eye movement with an angular velocity lower than 30 degrees/second is classified as a fixation, while all other points are classified as saccades [5]. Depending on factors such as the distance between the observer's eyes and the screen, as well as the screen's size and resolution, this angular velocity threshold is converted to a pixels/second threshold. The remaining gaze points correspond to the fixation points of the Sonographers.

2.2 Proposed Method

We proposed a scan-path based fine-tuning training mechanism(SFT), as shown in Fig. 2. Firstly, we propose the learnable kernel to simulate different visual scan search patterns instead of merely applying single gaze attention map directly as the supervision signals to train the CNN model. Secondly, in view of the lack of mutual adjustment between the generation of fixation graphs and diagnostic classification, we propose a joint optimization objective function of model prediction loss and scan-path interpretation loss.

Search Pattern Aware with Learnable Kernels. To recognize the sonographers' visual attention region or called search patterns, an intuitive way is to define $\boldsymbol{h}(\cdot)$ as applying a $k \times k$ Gaussian kernel on the gaze scan-path $\boldsymbol{G}$. Unfortunately, diagnosing gout from MSKUS using this method poses two challenges. Firstly, the generation of the gaze attention maps is independent of diagnosis

classification, with both processes lacking mutual adjustment and collaborative learning. Secondly, different sonographers have different eye scan search patterns, and diverse MSKUS have diverse feature information. This approach, which uses static methods to model each image individually, ignores the learning between the sonographers' eye scan pattern and the MSKUS feature information. Therefore, we further extend this idea by defining a learnable input function $\boldsymbol{h}_{\Phi}$ with multiple learnable kernel transformations so that the model can be more aware of the sonographer's visual attention area. The learnable imputation function $\boldsymbol{h}_{\Phi}$ can be realized by applying multiple layers of convolution operations with learnable kernels over the raw gaze scan-path $\boldsymbol{G}$. The weights of learnable kernels are constantly adjusted to the optimum during training.

Scan-path based Fine-tune Training Mechanism. Algorithm 1 explains our training mechanism in detail. Concretely, the proposed objective function is as follows:

$$\min_{\theta,\Phi} \sum_{i=1}^{N} \mathcal{L}_{predict}(\boldsymbol{f}_{\theta}(x^i), y^i) + \mathcal{L}_{scan-path}(\boldsymbol{h}_{\Phi}(\hat{\boldsymbol{G}}^i), \boldsymbol{M}^i) \tag{1}$$

For the MSKUS dataset $\boldsymbol{D} = \{x^i \in \mathbb{R}^{C\times H\times W},\ \ y^i \in \{0,1\}\}_{i=1}^{N}$, let x^i be the input image with C channels, H as height and W as width. Let y^i be the gout label of input x^i (with '1' to represent a gout patient and '0' for a healthy person, respectively). A CNN model learns the mapping function $\boldsymbol{f}_{\theta}$ for each input MSKUS image to its corresponding gout label $\boldsymbol{f}_{\theta} : x \rightarrow y$ and θ is the CNN model parameter. In addition, sonographers' visual scan-path is considered as a number of directed gaze points $\hat{\boldsymbol{G}}^i$ described in the previous subsection. Concretely, the $\boldsymbol{h}_{\Phi}$ optimization object in Equation. 1 involves optimizing both the parameter of learnable kernels Φ and the model-generated attention map $\boldsymbol{M}^i = g(\boldsymbol{f}_{\theta}({x_1}^i))$. In the equation, $g(\cdot)$ specify the CAM method [9], $\boldsymbol{M}^i \in \mathbb{R}^{H\times W}$ denotes CNN model-generated attention map for i-th sample; $\boldsymbol{h}_{\Phi}$ is the learnable kernel function discussed in the previous subsection and Φ is the parameter of learnable kernels function. We propose to optimize θ and Φ with a conventional gradient descent algorithm by proposing a differentiable approximation to the indicator function. Here, we only use the collected sonographers' visual scan-path in the model training phase, and do not require any additional input in the model testing phase.

3 Experiments

3.1 Experimental Settings

MSKUS Data Collection. Ethics approval for human data collection was obtained from Nanjing Drum Tower Hospital. All subjects provided written consent

Algorithm 1 Scan-path based fine-tune training mechanism

Input: Dataset $\boldsymbol{D}$, Pre-process scan-path $\hat{\boldsymbol{G}}^i$, Model parameters θ and Φ
Output: Optimized parameters θ^* and Φ^*
  for $epoch \leq N$ **do**
    for x^i, y^i in $\boldsymbol{D}$ **do**
      Model forward: $\tilde{y}^i = \boldsymbol{f}_\theta(x^i)$
      Calculate Predict loss: $\mathcal{L}_{predict} = CE(\tilde{y}^i, y^i)$
      Sonographers' visual attention map based on learnable kernels: $\boldsymbol{S}^i = \boldsymbol{h}_\Phi(\hat{\boldsymbol{G}}^i)$
      Calculate CNN model-generated attention map CAM: $\boldsymbol{M}^i = g(\boldsymbol{f}_\theta(x_1^i))$
      Calculate Scan-path loss: $\mathcal{L}_{scan-path} = MAE(\boldsymbol{M}^i, \boldsymbol{S}^i)$
      $Loss = \lambda_p \mathcal{L}_{predict} + \lambda_s \mathcal{L}_{scan-path}$
      Loss backward to optimization parameters θ^* and Φ^*
    end for
  end for

to participate in this study. The examinations were conducted using a Toshiba Aplio500 scanner (Toshiba, Tokyo, Japan) with a 5-12MHz linear array transducer. The joints in which gout lesions were detected included the bilateral knee, ankle, and first metatarsophalangeal joints of each patient. Dataset totally contains 1127 MSKUS images from different patients including 509 gout images and 618 healthy images. The resolution of the MSKUS images were resized to 224 $\times$ 224. During the experiments, we randomly divided 10% of the dataset into testing sets. Then we used 5-fold cross validation to divide the training sets and validation sets.

Evaluation Metrics. Five metrics were employed to quantitatively assess the performance of each model: Accuracy (ACC), Area Under Curve (AUC), Correlation Coefficient (CC), Similarity (SIM), and Kullback-Leibler divergence (KLD) [3]. ACC and AUC were utilized to evaluate the models' classification performance. CC, SIM, and KLD were used to ascertain the degree of alignment between the models' attention areas and those of sonographers during diagnoses.

Implementation Details We conducted the experiments using the Pytorch framework on a single NVIDIA GTX 2080TI GPU. The models were trained for 50 epochs using the ADAM optimizer with a learning rate of 1×10^{-4}. The batch size was set to 10.

3.2 Results

Comparison with SOTA Sonographer Attention-based Mechanism. For the gout classification task, as shown in Table. 1, we selected five classic CNN classification models (Complete comparison can be found in the appendix.) under different sonographer-attention based adjusting mechanisms including TLS mechanism [6] and our "Scan-path based fine-tune training mechanism" (SFT).

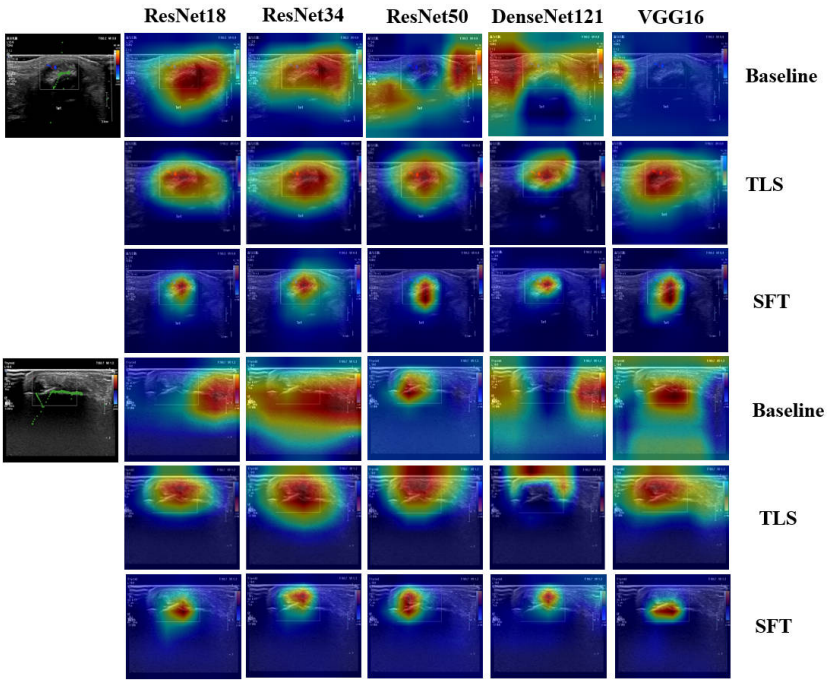

Fig. 3. Grad-CAM visualizations for the baseline, TLS and SFT across different backbone. Green points represents the scan-path from annotation procedure.

The results indicated that utilizing our SFT mechanism resulted in a significant enhancement in all metrics, compared to the TLS models. Specifically, our model's ACC and AUC metrics are slightly better than TLS models, highlighting its effectiveness in classification. The CC and SIM scores of our model were, on average, 0.175 and 0.038 higher than those of the TLS model, respectively. The KLD scores (Lower is better) exhibited a reduction of 0.498 in comparison to the TLS models. These results demonstrate that our SFT adjusting mechanism not only consistently improves the performance of CNN models in classifying gout, but also outperforms the existing SOTA TLS adjusting mechanism. Furthermore, Fig. 3 illustrated the qualitative results of CAM obtained by backbone CNN, CNN with TLS, and CNN with SFT. The backbone models without SFT had a tendency to focus on noise, textures, and artifacts, leading to unreliable gout diagnoses. Conversely, models with SFT were able to effectively concentrate on the critical lesion area, which aligned with the clinical diagnostic experience of sonographers.

Comparison with Other Human-attention-fused CNN Models. Some human-attention-fused CNN models incorporate recorded gaze information into deep CNN models, while also imposing specific network structure constraints.

Table 1. Comparison with other sonographer attention-based adjusting mechanism and fixed Gaussian kernel.

Models	ACC↑(%)	AUC↑	CC↑	SIM↑	KLD↓
ResNet50	83.385±5.705	0.924±0.028	0.247±0.025	0.247±0.025	2.168±0.063
ResNet50-TLS	89.617±3.055	0.967±0.011	0.402±0.028	0.298±0.020	2.133±0.232
ResNet50-Gaussian	86.305±3.344	0.937±0.025	0.530±0.048	0.329±0.034	1.517±0.102
ResNet50-SFT	90.615±2.142	0.974±0.008	**0.603±0.018**	0.367±0.015	**1.331±0.046**
DenseNet121	82.307±0.973	0.927±0.008	0.260±0.036	0.149±0.008	2.142±0.075
DenseNet121-TLS	89.454±1.621	0.965±0.010	0.369±0.011	0.239±0.007	1.991±0.062
DenseNet121-Gaussian	87.076±1.713	0.959±0.017	0.398±0.0497	0.222±0.014	1.791±0.089
DenseNet121-SFT	89.387±0.574	0.966±0.004	0.529±0.049	0.272±0.021	1.570±0.107
Vgg16	89.132±3.201	0.958±0.021	0.221±0.089	0.182±0.044	3.461±0.776
Vgg16-TLS	91.501±2.885	0.966±0.020	0.416±0.020	0.305±0.013	1.932±0.084
Vgg16-Gaussian	90.615±2.451	0.959±0.034	0.509±0.092	0.330±0.055	1.689±0.500
Vgg16-SFT	**91.534±2.525**	**0.976±0.010**	0.560±0.039	**0.376±0.040**	1.427±0.079

Table 2. Comparison with other human-attention-fused CNN models.

Models	ACC↑(%)	AUC↑	CC↑	SIM↑	KLD↓
Sono-net [4]	90.391±0.019	0.945±0.009	0.057±0.048	0.096±0.026	12.841±2.786
Resnet+Gaze [12]	88.354±0.018	0.942±0.014	0.371±0.009	0.269±0.006	1.936±0.128
Unet+Gaze [8]	85.676±0.020	0.928±0.014	-0.003±0.013	0.0762±0.007	6.347±1.032
ResNet50-SFT	**90.615±2.142**	**0.974±0.008**	**0.603±0.018**	**0.367±0.015**	**1.331±0.046**

These models [4,8,12] require the collection of eye movement data from sonographers for each image. As shown in Table. 2, in terms of model classification accuracy, our model surpasses the human-attention-fused model (The ACC reached as high as 90.615, and the AUC achieved a level of 0.974). In attention evaluation metrics, compared to the top-performing ResNet-Gaze model [12], the CC is higher by 0.232, SIM is higher by 0.098, and KLD is reduced by 0.605. Our training mechanism demonstrates a marked improvement over other human-attention-fused CNN models. What is worth mentioning further is that, in comparison to other human-attention-fused CNN models, no additional input is required for the testing phase of the model. Thus our model can bypass the need to collect the eye movement data of the sonographers during the classification of newly acquired MSKUS images.

Ablation Analysis. To evaluate the effectiveness of learnable kernels, we compared the gout diagnosis results of several classification models between learnable kernels (CNN-SFT) and fixed Gaussian kernels (CNN-Gaussian). Taking ResNet50 as an example, we can get following observation from Table. 1. 1) the Resnet50-Gaussian and our model, both applying joint optimization, outperform the baseline model on all performance metrics. This indicates that the mechanism can improve the model accuracy and explainability; 2) as the joint optimization and learnable kernel are added in sequence, the performance pro-

gressively improved, with ACC increasing from 86.305% to 90.615%, and SIM rising from 0.329 to 0.367, suggesting the superiority of our mechanism.

4 Conclusion

In this paper, we proposed a scan-path based fine-tune mechanism that instructs the model to incorporate sonographers' clinical diagnostic experience. This mechanism have three strategies: 1). the learnable kernel we proposed can be aware of the diverse scan patterns of sonographers. 2). The novel scan-path based fine-tune mechanism can mutually enhance gout classification and align the model of attention with the sonographers' scan-path patterns. 3). Extensive experiments show that Our method has better gout diagnostic performance and generalization ability, can be combined with different CNN backbone networks. Specially, Sonographers' annotations byproducts, used as direct training input, enable the model to operate without requiring any additional input during the test phase. Therefore, Our model has promising clinical application.

Acknowledgment. The authors acknowledge supports from National Nature Science Foundation of China Grants (82027807, U20A20389, 62271246), National Key Research and Development Program of China (2022YFC2405200), and Natural Science Foundation of Jiangsu Province (BK20221477).

Disclosure of Interests. The authors have no competing interests to declarethat are relevant to the content of this article.

References

1. Alsharid, M., Cai, Y., Sharma, H., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Gaze-assisted automatic captioning of fetal ultrasound videos using three-way multi-modal deep neural networks. Medical Image Analysis **82**, 102630 (2022)
2. Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D.: Sononet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound. IEEE Transactions on Medical Imaging **36**(11), 2204–2215 (2017). https://doi.org/10.1109/TMI.2017.2712367
3. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE transactions on pattern analysis and machine intelligence **41**(3), 740–757 (2018)
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. Cao, L., Yang, D., Wang, Q., Yu, Y., Wang, J., Rundensteiner, E.A.: Scalable distance-based outlier detection over high-volume data streams. In: 2014 IEEE 30th International Conference on Data Engineering. pp. 76–87 (2014). https://doi.org/10.1109/ICDE.2014.6816641

6. Cao, Z., Zhang, W., Chen, K., Zhao, D., Zhang, D., Liao, H., Chen, F.: Thinking like sonographers: A deep cnn model for diagnosing gout from musculoskeletal ultrasound. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 159–168. Springer (2023)
7. Han, S., Kang, H.K., Jeong, J.Y., Park, M.H., Kim, W., Bang, W.C., Seong, Y.K.: A deep learning framework for supporting the classification of breast lesions in ultrasound images. Physics in Medicine & Biology **62**(19), 7714 (2017)
8. Karargyris, A., Kashyap, S., Lourentzou, I., Wu, J.T., Sharma, A., Tong, M., Abedin, S., Beymer, D., Mukherjee, V., Krupinski, E.A., et al.: Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development. Scientific data **8**(1), 92 (2021)
9. 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)
10. Thiele, R., Schlesinger, N.: Diagnosis of gout by ultrasound (2007)
11. Voisin, S., Pinto, F., Xu, S., Morin-Ducote, G., Hudson, K., Tourassi, G.D.: Investigating the association of eye gaze pattern and diagnostic error in mammography. In: Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment. vol. 8673, p. 867302. SPIE (2013)
12. Wang, S., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Follow my eye: Using gaze to supervise computer-aided diagnosis. IEEE Transactions on Medical Imaging **41**(7), 1688–1698 (2022)
13. Zhang, Q., Gao, F., Sun, W., Ma, J., Cheng, L., Li, Z.: The diagnostic performance of musculoskeletal ultrasound in gout: a systematic review and meta-analysis. PLoS One **13**(7), e0199672 (2018)
14. Zhuang, Z., Yang, Z., Raj, A.N.J., Wei, C., Jin, P., Zhuang, S.: Breast ultrasound tumor image classification using image decomposition and fusion based on adaptive multi-model spatial feature fusion. Computer methods and programs in biomedicine **208**, 106221 (2021)