Longitudinal anatomical attention maps for recognizing diagnostic errors from radiologists' eye movements

Anna Anikina[1], Diliara Ibragimova[2], Tamerlan Mustafaev[3], Claudia Mello-Thoms[4], and Bulat Ibragimov[1]

[1] Department of Computer Science, University of Copenhagen, Denmark
{anan, bulat}@di.ku.dk
[2] Kazan State University Clinic, Kazan, Russia
[3] Swanson School of Engineering, University of Pittsburgh, Pittsburgh, United States
[4] Department of Radiology, University of Iowa, Iowa, United States

Abstract. With the rise in respiratory diseases, the workload on radiologists is increasing, leading to a higher risk of diagnostic errors. One approach to improve diagnostic processes is to reduce the frequency of cognitive and perceptual errors made by humans. This study aims to predict radiologists' diagnostic errors while interpreting chest X-rays using eye-tracking technology. We propose a novel method that combines human attention, derived from the locations of gaze fixation points, with attention from transformer neural networks. The resulting attention maps are combined with the segmentation of anatomical structures, including the lungs, clavicles, hila, heart, mediastinum, and esophagus, which restricts the analysis for regions potentially relevant for thoracic disease diagnosis. Attention maps are computed for each gaze fixation point, creating a longitudinal path representing the X-ray reading process. Finally, we applied Gated Recurrent Units (GRUs) to learn from the longitudinal attention maps and statistical gaze features to predict potential X-ray diagnostic errors. The proposed methodology was validated on $4,000$ chest X-ray readings performed by four radiologists. The model achieved an error detection accuracy of 0.79, measured as the area under the receiver operating characteristic (ROC) curve. The code is available at https://github.com/annshorn/TEGRU

Keywords: eye tracking · x-rays · visual transformer

1 Introduction

Despite significant advances in Artificial Intelligence (AI), integrating AI as a tool for computer-aided diagnosis (CAD) remains a challenging task; moreover, some patients are not ready to trust AI-generated diagnoses [1–4]. Therefore, instead of aiming for full automation and eliminating the human factor, AI can serve as an auxiliary tool in diagnostics. A promising approach for practical AI utilization into clinical workflows is to combine AI with eye-tracking sensors to assess radiologists' performance [5]. For example, such a combination allows models to reveal cognitive processes [6], understand the reading patterns of radiologists at different levels [7, 8], and assist in training of less experienced radiologists [9, 10]. It also helps detect radiologists' decision errors by directly capturing what radiologists look at and, more importantly, what they miss during image interpretation, as demonstrated in fields such as dentistry [11], mammography [12–14], and thoracic radiography [15–18]. While in traditional AI CAD, radiologists and AI complete their tasks separately, eye-tracking enables continuous communication between human and computer, allowing for quicker error detection and intervention [5].

Despite numerous studies, the field of predicting radiological errors from gaze data remains largely underinvestigated. Most studies focus on localized diseases like tumors, neglecting systemic conditions affecting multiple areas. Additionally, they analyze static gaze fixations without considering longitudinal gaze patterns. Such static analysis can sometimes turn into the estimation of whether the reader visually captured the abnormality location. In this work, we introduce a transformer-based algorithm that predicts radiological errors while reading chest X-rays with various abnormalities. The algorithm captured the reader's gaze over anatomically distinctive locations and analyzed the longitudinal patterns of such data. It also combined transformers with recurrent neural networks trained on statistical gaze features, learning from both longitudinal image and gaze information. The algorithm was validated on 4000 chest X-ray readings from four radiologists.

2 Methodology

2.1 Data collection

Experiment: Four board-certified radiologists ($3 - 30$ years of experience) analyzed 1000 publicly available chest X-rays from VinDr-CXR [19], CheXpert [20], RSNA [21], and SIIM-ACR [22] while their reading patterns were recorded. Among images, 420 did not contain any disease, 119 contained one disease, 143 contained two diseases, 318 contained three or more diseases, including pulmonary fibrosis (223 cases), aortic enlargement (193), cardiomegaly (152), pneumothorax (93), and other abnormalities like pleural thickening, lung opacity, pleural effusion, atelectasis, and nodules. To minimize fatigue and speed up diagnosis recording, radiologists provided detailed verbal reports instead of filling out paperwork. Each participated on two non-consecutive days without prior night shifts.
Technical details: To replicate typical working conditions, radiologists were placed in an isolated, quiet room with a 10-bit LG monitor (3840×2160 resolution, pixel density $\rho = 7.31$ px/mm), a Tobii Eye Tracker 4C (90 Hz), and a Logitech 960 microphone for voice recording. The eye tracker was positioned to avoid interfering with image reading. A specialized user interface was developed to minimize distractions, allowing control with a single "Enter" button. Diagnoses were dictated and manually processed for analysis.

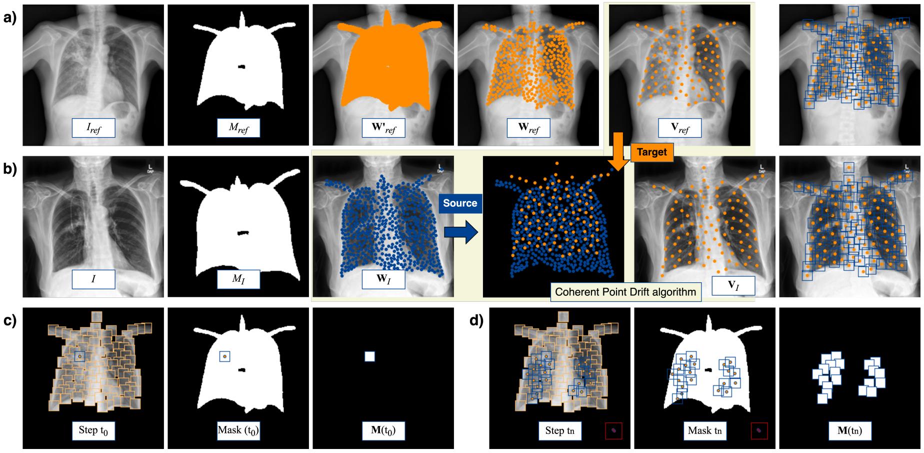

Fig. 1. Representation of image preparation for the Vision Transformer and formation of anatomical patches. Panel (a) shows the target X-ray image, whereas panel (b) shows the source X-ray image. The images are first segmented using a neural network. Both segmentations are uniformly filled with points, which then guide the Coherent Point Drift algorithm for aligning the target image anatomies to the source image anatomies. Finally, a subsample of registered points is selected as centers for anatomically aligned patches between the two images. Panel (c) and (d) show attention maps $\mathbf{M}$ construction.

2.2 Feature preprocessing

Calculation of fixation points: During the perception of visual information, two types of eye movements are involved: saccades and fixations. Fixation points are moments when the eye remains still and focused on a specific point in the scene. Saccades are characterized by rapid eye movements that occur between fixations. These two types of movements alternate with each other, but the information capture occurs only during fixations [23]. To identify fixation points, we calculated the angular velocity of the eye movements. Movements exceeding a threshold of 30 degrees per second were classified as saccades and excluded from further analysis [24].

Reading features: Previous studies show that fixation analysis reveals patterns linked to radiologists' experience [25, 26] and aids in evaluating diagnostic performance [13,27]. Based on this, we computed features such as fixation count, average distance and angle between fixations, and lung-to-lung transitions. To facilitate this, we introduced a distance matrix D_t, where each cell stores the distance between (x_t, y_t) to the nearest fixation point. We denoted P_t as a set of all fixation points that have been collected up to fixation t. Using distance matrix D_t, we calculated the coverage of the right and left lungs, and the amount of information the radiologist obtained at each new fixation. The visual coverage of the lungs for fixation points can be computed as $\psi(y,t) = \exp\left(-\frac{|(x,y)-y|^2}{2\cdot z\cdot \rho\cdot \tan^2\left(\frac{\theta}{2}\right)}\right)$,

where z is the distance between the monitor and the radiologist's eyes, θ is a viewing angle, and ρ is the pixel density of the monitor. The amount of new information about the lung area that the radiologist gains at each fixation point during image reading is calculated as $s(t) = \frac{\sum_y \psi(y,t) \cdot I_l(y)}{\sum_y I_l(y)}$, where I_l is the lung segmentation array obtained using a contour-aware U-Net [28], which outperforms standard U-Net by incorporating contour loss to penalize border segmentation errors, where mistakes are more frequent and harder to correct.
Longitudinal reading features: For each fixation point $(x, y) \in \mathbf{X}'$, where $\mathbf{X}' = \{(x_1, y_1), \ldots, (x_{\mathbf{T}'}, y_{\mathbf{T}'})\}$ defines the sequence and total reading length $\mathbf{T}'$, we computed d features as a vector: $(x_t, y_t) \mapsto \left[\zeta_1^{(t)} \ldots \zeta_d^{(t)}\right]$. The number of fixation points determines the feature set length. Since $\mathbf{X}'$ varies in $\mathbf{T}'$, we define a fixed $\mathbf{T}$: if $\mathbf{T}' < \mathbf{T}$, we pad; if $\mathbf{T}' > \mathbf{T}$, we merge initial fixations, precomputing their statistics in $\zeta_{1\ldots d}^{(t=t_0)}$. This ensures a uniform fixation set $\mathbf{X}$ and length $\mathbf{T}$.

2.3 Anatomical patches

Beyond gaze features, we analyzed which X-ray areas attracted radiologists' attention, as decision errors depend on both eye movement patterns and anatomical regions viewed. Effective analysis must consider fixation sequence, duration, and gaze returns while focusing only on relevant anatomical areas. To achieve this, we propose using anatomy- and gaze-restricted visual transformers (ViT). Unlike classical (grid-based) division images into uniform patches [29] for ViT, we propose to focus ViT on gaze over anatomically relevant areas. In particular, we used the pre-trained PSPNet from the TorchXRayVision library [30] to identify and segment the clavicles, lungs, hila, heart, aorta, mediastinum, and esophagus.

A randomly selected image I_{ref} and its segmentation masks M_{ref} (Fig. 1, the panel (a)) served as the reference anatomy. From the segmentation mask M_{ref} two sets of densely sampled $\mathbf{W}_{ref}$ and sparsely sampled $\mathbf{V}_{ref}$ points are extracted. For a previously unseen image I_n, and its automatically-generated segmentation masks M_{I_n} (Fig. 1, the panel (b)), the densely sampled set $\mathbf{W}_{I_n}$ was generated and then registered to $\mathbf{W}_{ref}$ using the Coherent Point Drift (CPD) algorithm. The transformation matrix of CPD allows us to deform sparsely sampled set $\mathbf{V}_{I_ref}$ towards I_n resulting in an anatomically consistent sparsely sampled set $\mathbf{V}_{I_n}$. CPD treated $\mathbf{V}_{ref}$ as Gaussian centers, optimizing their positions to maximize the likelihood of $\mathbf{V}_{I_n}$ by minimizing the negative log-likelihood function:

$$P(\mathbf{V}_{I_n} | \mathbf{V}_{ref}) = \prod_{i=1}^{N} \left(\sum_{j=1}^{M} w_j \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y}_i - \mathbf{x}_j\|^2 \right) + w_{out} \right), \qquad (1)$$

where w_j represents the weight of the j-th Gaussian component, indicating the probability that a data point y_i corresponds to the model point x_j; σ^2 is the variance of the Gaussians; and w_{out} accounts for the probability of outliers. Using point from sets $\mathbf{V}_I$ as centers, patches $\{p_l\}_{l=1}^{L}$ was formed, ensuring anatomical consistency across images. Finally, positional encoding from transformers [31] was applied to incorporate sequence information.

2.4 Attention maps

After obtaining $\{p_l\}_{l=1}^{L}$, we computed attention maps $\mathbf{M}$ (Fig. 1, the panel (c) and (d)), which guide the ViT to focus on relevant image regions by assigning varying weights. In other words, $\mathbf{M}$ should indicate which patches are important for calculating the attention score and which are not. Specifically, we incorporated the information about the reader's gaze into $\mathbf{M}$, i.e., visually observed areas up to the current fixation t. At each fixation t, the fixation point (x_t, y_t) is mapped to the image. If (x_t, y_t) falls within a segmented anatomical region, the corresponding patch p_l is marked as diagnostically important. Let $\mathcal{S}^{(t)}$ denote the set of patches that are considered important from the beginning up to fixation t, forming the attention map vector $\mathbf{M}^{(t)} = [\mathbf{M}_1^{(t)}, \mathbf{M}_2^{(t)}, \ldots, \mathbf{M}_L^{(t)}]$, where each element of $\mathbf{M}_l^{(t)}$ is defined as follows:

$$\mathbf{M}_l^{(t)} = \begin{cases} 0, & p_l \in \mathcal{S}^{(t)}, \\ -\infty, & p_l \notin \mathcal{S}^{(t)}, \end{cases} \tag{2}$$

for $\forall l \in \overline{1, L}$. The $\mathbf{M}^{(t)}$ is updated at the next fixation $t + 1$ based on the set $\mathcal{S}^{(t+1)}$, and this process continues throughout the sequence of fixations.

2.5 Network for diagnostic error prediction

The model receives two input sequences corresponding to reading features and visual X-ray information. The first sequence has dimensions of $(\mathbf{T} \times \zeta)$, where $\zeta = \left[\zeta_i^{(t)}\right]_{t=1:T, i=1:d}$ represents the reading features matrix. The second sequence contains $\{p_l\}_{l=1}^{L}$ with dimensions $(L \times w \times h)$, where w and h are the width and height of a patch. For each fixation t, the map $\mathbf{M}^{(t)} = [\mathbf{M}_1^{(t)}, \mathbf{M}_2^{(t)}, \ldots, \mathbf{M}_L^{(t)}]$ is generated based on $\mathcal{S}^{(t)}$.

Transformer Encoder: The ViT's input consists of $\{p_l\}_{l=1}^{L}$ and $\mathbf{M}^{(t)}$. Each patch p_l is linearly transformed into embedding ϵ_l. At the beginning of the input embeddings $\epsilon = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_L\}$, we add a classification (CLS) token $\epsilon_{CLS}^{(t)}$. The role of the CLS token is to act as an aggregate representation of the entire image, summarizing the information after it has been processed through the transformer layers. We modified the mask $\mathbf{M}^{(t)}$ to fit the requirements of the multi-head attention (MHA) block, where each of h heads performs a separate attention function. The mask $\mathbf{M}^{(t)}$ was duplicated for each head and then expanded to the size $L \times L$ by doubling the elements for each patch length. For each head $i, i \in \overline{1, h}$, the input ϵ is linearly projected to queries $\mathbf{Q_i}$, keys $\mathbf{K_i}$, and values $\mathbf{V_i}$. The attention score $\mathrm{A}(\mathbf{Q_i}, \mathbf{K_i}, \mathbf{V_i})_t$ is then calculated and concatenated into a single matrix $\mathrm{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V})_t = \mathrm{A}(\mathbf{Q_1}, \mathbf{K_1}, \mathbf{V_1}) \oplus \cdots \oplus \mathrm{A}(\mathbf{Q_h}, \mathbf{K_h}, \mathbf{V_h})$ [29].

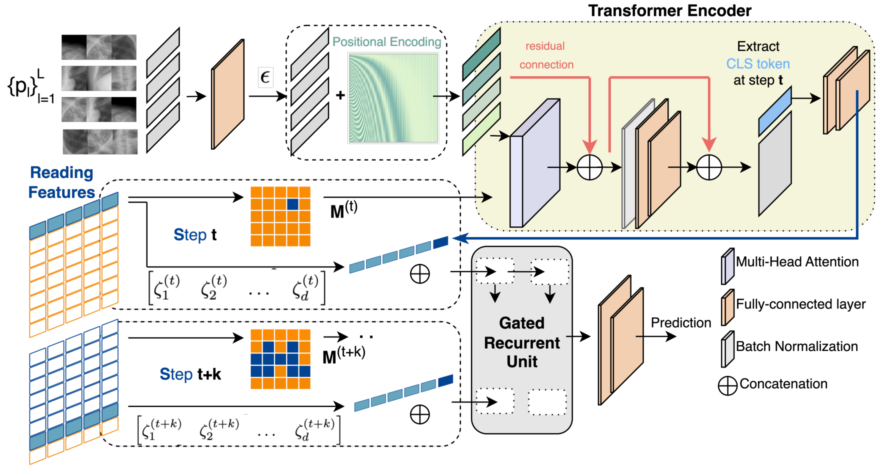

Fig. 2. Illustration of the proposed neural network. As an initial step, the input image is decomposed into anatomical patches $\{p_l\}_{l=1}^{L}$ and longitudinal reading features. At each fixation t, the transformer encoder first receives all the patches $\{p_l\}_{l=1}^{L}$ and a map $\mathbf{M}^{(t)}$. This map indicates which areas of the image the radiologist has already reviewed and are therefore considered important. The vector representation of the patches $\{p_l\}_{l=1}^{L}$ passes through a linear layer to obtain embeddings ϵ. Then, a classification (CLS) token is added, and positional encoding is applied. The ϵ along with $\mathbf{M}^{(t)}$ pass through the Transformer Encoder to generate the current state embedding. The extracted CLS token is combined with the reading features $\left[\zeta^{(\mathbf{t})} \otimes \epsilon_{CLS}^{(t)}\right]$ and fed into a Gated Recurrent Unit (GRU), which are then used to predict the reading error as a weighted sum of the GRU outputs.

Following with, with a fully connected layer, we extract the classifier token $\epsilon_{CLS}^{(t)}$ from the transformer encoder.

Prediction with Gated Recurrent Unit (GRU): The $\epsilon_{CLS}^{(t)}$ is concatenated with $\zeta^{(\mathbf{t})} = \left[\zeta_1^{(t)}\ \zeta_2^{(t)} \ldots \zeta_d^{(t)}\right]$ to form an input vector to the GRU cell. At each fixation t, the GRU output h_t is generated and concatenated into a matrix τ of size $\mathbf{T} \times H_{out}$. The τ passes through a multilayer perceptron (MLP) to calculate the predictions of the X-ray reading outcome for each step from 1 to $\mathbf{T}$. Considering that we are much less certain about the reading outcome when only the first few fixations are observed, compared to when most of the fixations are observed, different weights should be assigned to prediction at different fixation points. Following this logic, we weight each GRU outcome using coefficient vector $\boldsymbol{w} = \left[\frac{2k}{\mathbf{T}(\mathbf{T}+1)}\right]_{k=1}^{\mathbf{T}}$ and calculate the predicted X-ray reading error as the weighted sum of the outcomes.

Table 1. Results of the comparison between existing and proposed methods on the collected dataset. [(1)] Spatial Frequency Bands (SFB) are the spatial frequency bands determined based on the wavelet coefficients obtained by decomposing the image using the stationary wavelet packet transform at the 3d level of decomposition. [(2)] Region of Interest (ROI) is a specific area within an image that attracts attention or contains abnormality. [(3)] Statistical and Gabor features represent first-order and second-order statistics (entropy, standard deviations, contrast, correlation, energy, homogeneity) and Gabor wavelet features.

Method	Longitudinal features	Transformer based	AUC	Micro Average			Macro Average		
				Precision	Recall	F1-score	Precision	Recall	F1-score
Proposed with Anatomical patches	✓	✓	**0.7869**	**0.7463**	**0.7463**	**0.7463**	**0.7411**	**0.7352**	**0.7373**
ResNetAE w/c GRU [18]	✓		0.7755	0.7266	0.7266	0.7266	0.7266	0.732	0.725
TransGATConv [17]	✓	✓	0.7442	0.6798	0.6798	0.6798	0.6886	0.6914	0.6795
Proposed with classical (grid-based) patches	✓	✓	0.7228	0.6675	0.6675	0.6675	0.6948	0.6892	0.667
SVC with SFB(1) features [15]			0.6792	0.6595	0.6595	0.6595	0.6535	0.6589	0.6535
ResNet-152 with ROI(2) [14]			0.614	0.5663	0.5663	0.5663	0.5919	0.5911	0.5662
J48 with Statistical and Gabor features(3) [16]			0.5728	0.617	0.617	0.617	0.591	0.5733	0.5684
iALD (base ResNet-152) [32]			0.5263	0.5334	0.5334	0.5334	0.5396	0.5299	0.5012

3 Experiment and results

Database: From $4,000$ X-ray reading, we removed readings that were too short, i.e. contained only a few fixations or had many corrupted recording points when the reader moved far away from the screen. As a result, the training set included 826 cases ($2,644$ readings with correct diagnoses, and 612 with diagnostic errors), and testing contained 134 cases (234 correct and 173 incorrect diagnoses). Since some X-rays contained multiple abnormalities, a diagnostic error was defined as a case where the radiologist either did not mention any of the abnormalities present or mentioned an abnormality while the X-ray contained none.

Training: The model input included a feature sequence $(\mathbf{T} \times \zeta) = (97 \times 27$ and patches of size $(L \times w \times h) = (97 \times 16 \times 16)$. If multiple points (x_1, y_1) and (x_2, y_2) fell within the same patch, it was added to $\mathcal{S}$ only once, with fixation statistics updated accordingly. Each patch was represented as a 256-length vector, transformed via a linear layer into embeddings ϵ of size (97×512). By adding the CLS token, the dimension became (98×512). The MHA used 2 heads with a dropout of 0.1. The modified mask $\mathbf{M}^{(t)}$ had a size of $(2 \times 97 \times 97)$. After passing through the transformer encoder, a vector $\epsilon^{(t)}_{CLS}$ of length 16 was extracted and concatenated with $\zeta^{(\mathbf{t})}$, forming length of 43. The GRU had 6 layers with a h_t of 16, producing an output matrix τ of size $(\mathbf{T} \times H_{out}) = (97 \times 16)$. A final linear layer with softmax generated predictions for 97 steps. The model was trained with a batch size of 8 for 100 epochs using Adam optimizer at a learning rate of 1×10^{-5}.

Results: Table 1 presents the results of comparing the proposed method with existing methods dedicated to recognizing diagnostic errors by radiologists. The results indicate that longitudinal features outperform other feature-based methods in terms of metrics. It is important to emphasize that while most existing methods focus on localized diseases [14–16], we are testing these methods on both localized diseases and diseases that affect multiple areas. For each method, we followed the exact protocol suggested by the authors by making some adjustments to accommodate the algorithm to our data format.

Ablation experiments: Table 1 confirms that the proposed use of anatomical patches outperforms classical (grid-based) patches. The use of CPD ensures the correspondence between the patches, i.e., the $l-$th patch p_l^m on image I_m anatomically corresponds to the $l-$th patch p_l^n on image I_n. With this property, the model is expected to be more resistant to changes in input images, such as rotations. To test this, we applied a random 10° rotation and 10% scaling to test X-rays and gaze data and measured the resulting changes in the error prediction performance. Accuracy declined by only $0.6 - 0.9\%$ AUC with CPD-based patches but dropped significantly by $3.43 - 4\%$ AUC for standard grid-based patches.

4 Discussion

In this work, we explored the application of eye-tracking data to predict diagnostic errors made by radiologists during chest X-ray readings. Methodologically, we proposed augmenting the transformer neural network with anatomically sampled patches and an attention matrix corresponding to the fixation-based attention. Such an approach bridged the intuition about human perception and image processing with neural networks. Moreover, it allowed us to exclude from consideration visual attention focused on screen areas outside the target anatomical structures, which pollutes the gaze data.

Patch size is a hyperparameter, with studies exploring its optimal value for transformer training. Nguyen et al. [33] found that smaller patches improve transformer results, peaking at 1×1 pixel-size patches. However, transformer complexity grows non-linearly as patch size decreases, since sequence length increases proportionally to the inverse square of patch size [29]. Smaller patches also often process less informative areas (e.g., background). Anatomical patches avoid this issue by focusing on relevant regions. In our case, the patch size should be small enough to improve transformer performance but large enough to capture meaningful anatomical information. A 16×16 patch corresponds to a viewing angle of 1.73°, close to the foveal vision range (1.5°) [34]. Using anatomical patches instead of classical (grid-based) equidistant patches also reduced computational complexity by focusing only on relevant image regions.

We trained a RandomForestClassifier only on gaze features, achieving an AUC of 0.694, outperforming non-longitudinal features (Table 1). This observation indicates that radiologists' way of reading X-rays changes when they make errors. We computed the odds ratio (ORs) of the erroneous diagnosis if an individual gaze feature was above or below the median value for this feature for all participating radiologists. The number of gaze switches between lung fields, i.e., moments when fixation moves from left to right lung field or vice versa, had an OR of 0.4, meaning that the readings with a high number of switches are

2.5 times more likely to be erroneous than the readings with a low number of switches. Features that also exhibited statistically significant ORs included the number of fixations per reading (OR = 0.44), number of visits of the left (OR = 0.48) and right (OR = 0.48) lung fields, and mean information gain (OR = 2.1). High mean information gain indicates that the reader gains a lot of new information from each fixation, i.e., they cover the image effectively without getting stuck around already observed areas. Overall, the readings that result in a correct diagnosis are more likely to have fewer fixations, which are distributed more efficiently over the anatomy. This is in agreement with observations in eye-tracking literature [35].

Acknowledgments. This work was supported by the Novo Nordisk Foundation under Grant NFF20OC0062056, and the National Institutes of Health under grant 1R01CA259048.

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

References

1. Handa T. The potential role of artificial intelligence in the clinical practice of interstitial lung disease. *Respiratory Investigation*, 61(6):702–710, 2023.
2. Yin J. et al. Role of artificial intelligence applications in real-life clinical practice: Systematic review. *Journal of Medical Internet Research*, 23(4):e25759, 2021.
3. Ibragimov B., Arzamasov K., et al. A 178-clinical-center experiment of integrating ai solutions for lung pathology diagnosis. *Scientific Reports*, 13(1):1135, 2023.
4. Han R. et al. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *The Lancet Digital Health*, 6(5):e367–373, 2024.
5. Ibragimov B. and Mello-Thoms C. The use of machine learning in eye tracking studies in medical imaging: A review. *IEEE Journal of Biomedical and Health Informatics*, 28(6):3597–3612, 2024.
6. Brunyé T., Nallamothu B., and Elmore J. Eye-tracking for assessing medical image interpretation: A pilot feasibility study comparing novice vs expert cardiologists. *Perspectives on Medical Education*, 8, 2019.
7. Castner N., Kuebler T.C., et al. Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing. In *ACM Symposium on Eye Tracking Research and Applications*, ETRA '20 Full Papers.
8. Ahmidi N., Ishii M., et al. An objective and automated method for assessing surgical skill in endoscopic sinus surgery using eye-tracking and tool-motion data. *International Forum of Allergy & Rhinology*, 2(6):507–515, 2012.
9. Ahmidi N. et al. Surgical task and skill classification from eye tracking and tool motion in minimally invasive surgery. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI*, volume 13, Part 3, pages 295–302, 2010.
10. Chetwood A.S.A., Kwok K.W., et al. Collaborative eye tracking: a potential training tool in laparoscopic surgery. *Surgical Endoscopy*, 26:2003–2009, 2012.
11. Castner N., Klepper S., et al. Overlooking: The nature of gaze behavior and anomaly detection in expert dentists. 10 2018.

12. Mello-Thoms C. Perception of breast cancer: Eye-position analysis of mammogram interpretation. *Academic Radiology*, 10(1):4–12, 2003.
13. Voisin S. et al. Predicting diagnostic error in radiology via eye-tracking and image analytics: Preliminary investigation in mammography. *Medical Physics*, 40, 2013.
14. Mall S., Krupinski E., and 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*, volume 10952, page 1095216, 2019.
15. Pietrzyk M., Donovan T., et al. Classification of radiological errors in chest radiographs, using support vector machine on the spatial frequency features of false-negative and false-positive regions. *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, 7966, 2011.
16. Tourassi G., Voisin S., et al. Investigating the link between radiologists' gaze, diagnostic decision, and image content. *Journal of the American Medical Informatics Association*, 20(6):1067–1075, Nov-Dec 2013. Epub 2013 Jun 20.
17. Anikina A., Karimzadeh R., et al. Prediction of radiological diagnostic errors from eye tracking data using graph neural networks and gaze-guided transformers. In *Graphs in Biomedical Image Analysis*, pages 33–42, 2025.
18. Anikina A. et al. Prediction of radiological decision errors from longitudinal analysis of gaze and image features. *Artificial Intelligence in Medicine*, 2024.
19. Nguyen H.T., Lam K.N., et al. Vindr-cxr: An open dataset of chest x-rays with radiologist annotations. *Scientific Data*, 9(1):1–10, 2022.
20. Irvin J., Rajpurkar P., et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:590–597, 07 2019.
21. Wang X. et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
22. Filice R.W. et al. Crowdsourcing pneumothorax annotations using machine learning annotations on the nih chest x-ray dataset. *Journal of Digital Imaging*, 2020.
23. Burr D.C., Morrone M.C., et al. Selective suppression of the magnocellular visual pathway during saccadic eye movements. *Nature*, 371(6497):511–513, Oct 1994.
24. Clare T., Lee L., et al. Skill characterisation of sonographer gaze patterns during second trimester clinical fetal ultrasounds using time curves. 2022, Jun 2022.
25. Kelly B.S. et al. The development of expertise in radiology: In chest radiograph interpretation 'expert' search pattern may predate 'expert' levels of diagnostic accuracy for pneumothorax identification. *Radiology*, 280(1):252–260, Jul 2016.
26. Tien T. et al. Differences in gaze behaviour of expert and junior surgeons performing open inguinal hernia repair. *Surgical Endoscopy*, 29(2):405–413, 2015.
27. Mugglestone M.D. et al. Defining the perceptual processes involved with mammographic diagnostic errors. In *Medical Imaging 1996: Image Perception*, volume 2712, pages 71 – 77. International Society for Optics and Photonics, SPIE, 1996.
28. Kholiavchenko M. et al. Contour-aware multi-label chest x-ray organ segmentation. *International Journal of Computer Assisted Radiology and Surgery*, 15, 2020.
29. Dosovitskiy A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.
30. Cohen J.P., Viviano J.D., et al. TorchXRayVision: A library of chest X-ray datasets and models. In *Medical Imaging with Deep Learning*, 2022.
31. Vaswani A., Shazeer N., et al. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

32. Mall S., Brennan P.C., and Mello-Thoms C. Can a machine learn from radiologists' visual search behaviour and their interpretation of mammograms—a deep-learning study. *Journal of Digital Imaging*, 32:746–760, 2019.
33. Nguyen D.K., Assran M., et al. An image is worth more than 16x16 patches: Exploring transformers on individual pixels. In *The Thirteenth International Conference on Learning Representations*, 2025.
34. Hans S., Ingo R., and Martin J. Peripheral vision and pattern recognition: A review. *Journal of Vision*, 11(5):1–84, 2011.
35. Brunyé T.T. et al. A review of eye tracking for understanding and improving diagnostic interpretation. *Cognitive Research: Principles and Implications*, 4, 2019.