Artificial Intelligence for the Analysis of Workload-Related Changes in Radiologists' Gaze Patterns

Journal:	*IEEE Journal of Biomedical and Health Informatics*
Manuscript ID	JBHI-00479-2022.R1
Manuscript Type:	Paper
Date Submitted by the Author:	06-May-2022
Complete List of Authors:	Pershin, Ilya; Innopolis University Kholiavchenko, Maksim; Rensselaer Polytechnic Institute Maksudov, Bulat; Dublin City University Institute of Education Mustafaev, Tamerlan; Public Hospital #2, Department of Radiology; Innopolis University Ibragimova, Dilyara; University Clinic of Kazan State University Ibragimov, Bulat; University of Copenhagen, Computer Science

Artificial Intelligence for the Analysis of Workload-Related Changes in Radiologists' Gaze Patterns

Ilya Pershin, Maksim Kholiavchenko, Bulat Maksudov, Tamerlan Mustafaev, Dilyara Ibragimova, Bulat Ibragimov*

***Abstract*—Around 60-80% of radiological errors are attributed to overlooked abnormalities, the rate of which increases at the end of work shifts. In this study, we run an experiment to investigate if artificial intelligence (AI) can assist in detecting radiologists' gaze patterns that correlate with fatigue. A retrospective database of lung X-ray images with the reference diagnoses was used. The X-ray images were acquired from 400 subjects with a mean age of 49 ± 17, and 61% men. Four practicing radiologists read these images while their eye movements were recorded. The radiologists passed a series of concentration tests at prearranged breaks of the experiment. A U-Net neural network was adapted to annotate lung anatomy on X-rays and calculate coverage and information gain features from the radiologists' eye movements over lung fields. The lung coverage, information gain, and eye tracker-based features were compared with the cumulative work done (CDW) label for each radiologist. The gaze-traveled distance, X-ray coverage, and lung coverage statistically significantly ($p < 0.01$) deteriorated with cumulative work done (CWD) for three out of four radiologists. The reading time and information gain over lungs statistically significantly deteriorated for all four radiologists. We discovered a novel AI-based metric blending reading time, speed, and organ coverage, which can be used to predict changes in the fatigue-related image reading patterns.**

***Index Terms*—eye-tracking, artificial intelligence, lung fields, chest X-ray, Unet, image segmentation**

I. Introduction

The workload of radiologists has been steadily increasing in the last decade mainly due to the increasing number of medical images acquired for diagnostic, treatment planning, and post-treatment monitoring purposes [1]. Artificial intelligence (AI) seems to be the tool that will significantly aid radiologists, especially in the field of computer-assisted diagnosis [2]. At the same time, the practical AI integration can rarely eliminate humans from the loop, and radiologists are usually responsible for the final diagnostic decisions [3]. The quality of human decisions is, however, far from being perfect with up to 4-10% of erroneous medical image readings [4]. Around 60-80% of radiological errors belong to perceptual errors, where abnormalities are overlooked, while the rest are cognitive errors, where abnormalities are seen but misinterpreted [5]. The radiological errors are influenced by different factors including the a priori expectations [6], visual bias [7], experience [8], and fatigue [9]. Understanding the effects of fatigue on the diagnosis and discovering the means for radiologists' fatigue detection has been of significant clinical interest [10].

Being a subjective feeling of overall tiredness, fatigue can be estimated using questionaries, e.g., Swedish Occupational Fatigue Inventory (SOFI) and Simulator Sickness Questionnaire (SSQ). The questionaries confirmed high fatigue levels after night shifts [11] and captured a faster fatigue growth for less experienced radiologists [12]. Several studies investigated how subjective questionaries can be replaced with objective tests. It was demonstrated that fatigue of physicians correlated with oxygenated hemoglobin concentrations in the prefrontal cortex [13] and electroencephalography changes [14]. Although more reliable than questionaries, such tests cannot be installed at radiologists' workstations due to high equipment requirements.

Eye movement analysis was shown to be a reliable and easy-to-deploy solution for radiologists' performance evaluation [15]. Eye-tracking helped discover two main strategies for medical image reading: bottom-up or stimulus-driven approach and top-down or knowledge-based systematic scanning approach [16]. It was observed that radiologists tend to spot

We thank Khanov A.N. MD and Zinnurov A.R. MD for participating in the experiment. This work has been supported by the Russian Science Foundation under grant #18-71-10072.

I. Pershin is with Innopolis University, Innopolis, Russia (e-mail: i.pershin@innopolis.ru)

M. Kholiavchenko is with Rensselaer Polytechnic Institute Troy, NY, USA (e-mail: kholim@rpi.edu). He was with Innopolis University, Innopolis, Russia.

B. Maksudov is with Dublin City University, Dublin, Ireland (e-mail: bulat.maksudov2@mail.dcu.ie). He was Innopolis University, Innopolis, Russia.

T. Mustafaev is with Innopolis University, Innopolis, Russia and University Clinic, Kazan State University, Kazan, Russia (e-mail: t.mustafaev@innopolis.ru)

D. Ibragimova is with the University Clinic, Kazan State University, Kazan, Russia (e-mail: ibragimovda@mail.ru)

B. Ibragimov is with the University of Copenhagen, Copenhagen, Denmark (e-mail: bulat@di.ku.dk).

* Indicates the corresponding author

regions with abnormalities in the first several seconds of reading, then deeper investigate such regions, and finally perform a comprehensive image scanning [17]. The expert read images more efficiently than novices and trainees by finding the areas of interest faster [18], [19] and paying more attention to such regions [20]. Eye-tracking can be used to assist with medical image annotation – a process that is usually done manually and therefore expensive and time-consuming [21]. Alternatively, radiologists' gaze movement patterns can be analyzed by machine learning algorithms to improve automated disease diagnosis and abnormality annotation [22], [23]. Fatigue level estimation from eye movements of radiologists has been also a topic of interest [24]–[26]. The existing studies showed the correlation between fatigue and pupil dilation and eyeblink rate [27], [28]. The gaze patterns over the target images seem to change with the fatigue growth, which has been observed by researchers but not always reflected in numeric metrics such as reading time, gaze traveled distance, number of fixation points, etc. [29].

Our paper, to the best of our knowledge, presents the first attempt to incorporate machine learning methods into the analysis of radiologists' gaze patterns during medical image examination to estimate their fatigue levels. The core idea is to apply deep neural networks for the segmentation of lung fields from chest X-rays and use the obtained segmentations to automatically measure the lung coverage with radiologists' gaze. The study hypothesis is that the gaze pattern will change with fatigue growth and that the proposed automated analysis can be used to reliably capture such changes. Four practicing radiologists were recruited to read chest X-rays while their eye movements were recorded. Four concentration tests were conducted after certain experiment milestones to check radiologists' fatigue/concentration loss. The statistical patterns in lung coverage other eye-movement features and concentration test results were evaluated against the reference metric of cumulative work done (CWD).

II. Methodology

A. Database

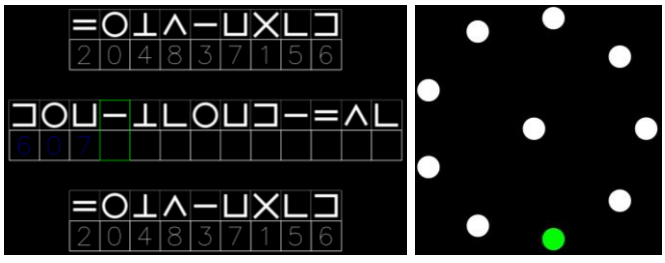
A public database VinDr-CXR, consisting of 18,000 posteroanterior chest X-rays collected from two hospitals in Vietnam, was used in this experiment [30]. The images were acquired by different modern scanners and were of different resolutions ranging from 1624×1775 to 3320×3408 pixels. Each image in the database was annotated by three radiologists from Vietnam hospitals. The annotations were defined as bounding boxes encompassing image areas with chest abnormality manifestations. In total, 27 abnormality types were labeled. An image was considered healthy if radiologists did not place any bounding box. We randomly sampled 400 X-rays from this database for this experiment. Among 400 sampled X-rays, 168 images were from healthy subjects, 60 had nodules/masses, 72 had infiltrations, 48 had pneumothorax, 40 had cardiomegaly, and 12 had atelectasis. For a pathological image to be sampled, all three radiologists should agree with the diagnosis. The proportion of healthy/pathological cases was justified by the distribution of pathologies in the original database [30] and by the desire to keep each pathology sufficiently present in the data. The abnormality composition and demographic information available for some of the cases are summarized in Table 1. Note that one patient could have multiple abnormalities or/and one abnormality can be presented in multiple places.

TABLE I
THE SUMMARY OF THE EXPERIMENT DATABASE WITH THE REFERENCE ABNORMALITY REPORTS.

Characteristic	Value
Patients	400
Demographics	
Gender (available), male:female	61%:39%
Age, mean and range	49 [4-90]
Chest abnormality manifestations	
No finding	168
Cardiomegaly	62
Aortic enlargement	53
Pleural thickening	82
Fibrosis	99
Nodule/Mass	76
Opacity	102
Atelectasis	24
Infiltration	80
Effusion	66
Calcification	11
Consolidation	44
Pneumothorax	49
Other lesions	53
# X-rays with	
1 abnormality	52
2 abnormalities	35
3 abnormalities	51
4 abnormalities	23
5 abnormalities	27
> 5 abnormalities	44

B. Experiment Setup

Four practicing radiologists were recruited to participate in the experiment. Radiologist A reads both X-ray and CT image modalities in his everyday clinical work. Radiologist B specializes in reading CT images, while Radiologists C and D exclusively read the X-ray image modality in their clinical work. The clinical experience was 3, 4, 3, and more than 30 years for Radiologists A, B, C, and D, respectively. The radiologists were unaware of the experiment aims so they were not tempted to change their image reading behavior. They were only informed that we want to analyze their diagnostic performance, record their eye movements and ask them to pass some concentration/reaction tests during the experiment [31]–



(a) (b)

Figure 1. Graphical user interfaces for the (a) digit symbol substitution (DSST) and (b) reaction time (RTT) tests. These tests were executed at specific time points of the eye tracking experiment to estimate the level of concentration and reaction of the participating radiologists.

[33]. During a one-day session, each radiologist analyzed all 400 X-rays separated into four batches of 100 X-rays, where the proportions of abnormalities and healthy cases matched the proportions in the complete database. The batches and X-rays inside them were individually shuffled for each radiologist.

A radiologist workstation was assembled in an isolated room at our institution. The workstation was equipped with an LG diagnostic 10-bit monitor with a resolution of 3840×2160 pixels and a pixel density of 7.21 px/mm, Tobii Eye Tracker 4C, and a microphone for voice recording. An in-house developed framework for X-ray image analysis was installed on the workstation. The framework was designed to minimize the use of the keyboard and mouse so that the user's gaze will be minimally distracted from the screen. Controller commands, eye tracking, and voice timestamps were recorded and synchronized.

Every radiologist was instructed to follow a predefined protocol to standardize the procedure and minimize eye movements unrelated to the X-ray reading. Starting, pausing, and finishing the reading of each X-ray was controlled by the Enter button. The X-ray contrast and brightness were adjustable using the mouse if needed. To avoid unnecessary typing, the radiologists verbally articulated the decision-making process, final diagnosis, and confidence level for each image. The information was manually extracted by a human listener from the voice recordings.

C. Experiment Execution

For each radiologist, the experiment happened on a weekend, after a full night's sleep. Before starting the experiment, a radiologist went through a 10-minute-tutorial to get familiar with the framework using example chest X-rays. The radiologist was asked to adjust the height and inclination of the chair to maximize working comfort. A technical assistant ensured that the radiologist's position is inside the recommended range of the eye-tracking, which lies between 20 to 37 inches. The distance between the user and eye-tracking has been automatically monitored by the framework, and a specific sound warned the user if he moves to the margins of the eye-tracker operation range.

The X-rays were analyzed in batches separated by fatigue measurements, eye-tracker calibration, and short breaks. After the analysis of the first two batches, each radiologist had a 40-minute lunch break. The level of fatigue was also measured at the start and end of the experiment. Due to software issues, the fatigue measurements were not correctly recorded for Radiologist A. As X-rays can be only shown once to each radiologist, there was no option to re-run the experiment for Radiologist A. The fatigue measurements were therefore analyzed only for Radiologists B-D.

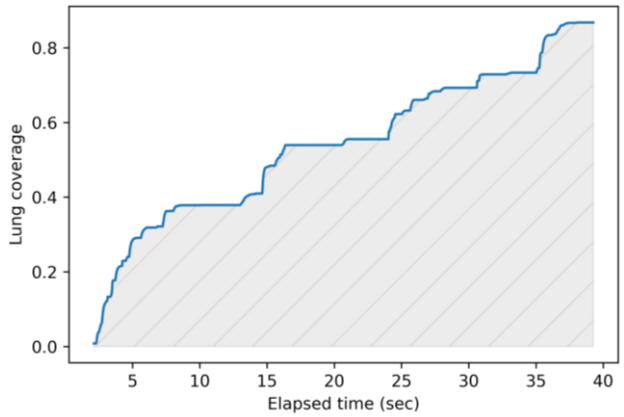


Figure 2. An example of the information gain histogram.

D. Fatigue/concentration measurements

We implemented four fatigue/concentration measurement tests. The first test was based on fatigue self-evaluation with SSQ [31]. This test is a questionnaire that consists of 16 questions grouped into oculomotor strain, disorientation, and nausea subscales. Although SSQ is usually used in virtual reality studies, its oculomotor subscale has been shown applicable for fatigue estimation in radiology. For example, Krupinski et al. [31] and Ikushima et al. [12] have demonstrated that the oculomotor subscale of SSQ predicts radiologists' fatigue and significantly correlates with the drop in diagnostic accuracy after a day of clinical reading. The oculomotor subscale of SSQ includes the following questions: general discomfort, fatigue, headache, eye strain, difficulty focusing and concentrating, and blurred vision. Each question is assessed by the participant on a four-point scale. Thus, the total test score is the sum of the points for all questions. A higher score corresponds to greater oculomotor fatigue. In the following parts of the paper, references to SSQ correspond to the oculomotor subscale of SSQ.

The second evaluation was performed using the digit symbol substitution test (DSST) [32]. The participant was asked to rapidly recover the sequence of digits encoded with graphical symbols using a look-up table with digit-symbol correspondences (Fig. 1a). The performance metric on DSST was based on the substitution accuracy and elapsed time.

The third evaluation was performed using a modified circle coverage test [33]. A circle with radius r_{circle} was shown in the center of the workstation screen. The participant was asked to visually traverse the circle contour in a rapid and maximally accurate manner. The evaluation metric was defined as a mean gaze trajectory deviation normalized to the elapsed time T:

$$DSST = \frac{\sum_{t=0}^{T}(|p_t| - r_{circle})^2}{T} \quad (1)$$

where p_t is the gaze coordinate at time point t, normalized against the circle center (Fig. 1b).

The fourth evaluation was performed using a reaction time test (RTT). Ten circles were arranged at the corners and center of a nonagon shown on the workstation screen. At the start of the test, the central circle was colored green and the participant was asked to focus the gaze on this circle. Every 3 seconds, the currently green circle turned gray and a new random circle turned green. The participant was instructed to move his gaze to the green circles as fast as he/she can. The test measured the average time needed for the participant's gaze to reach green circles.

E. Deep learning-based gaze coverage analysis

1) Lung Segmentation

Although highly personalized, chest X-ray reading usually follows specific clinical guidelines. These guidelines are based

TABLE II
THE CORRELATIONS BETWEEN FEATURES THAT CHARACTERIZE X-RAY IMAGE READINGS AND THE CUMULATIVE WORK DONE (CWD) BY EACH RADIOLOGISTS. THE FEATURES ARE CALCULATED USING THE RADIOLOGISTS' GAZE INFORMATION AND/OR DEEP LEARNING-BASED X-RAY IMAGE SEGMENTATION. EACH TABLE CELL CONTAINS A PERSON CORRELATION COEFFICIENT WITH 95% CONFIDENCE INTERVAL. STATISTICALLY SIGNIFICANT CORRELATIONS ARE HIGHLIGHTED IN GREEN.

Feature	Radiologist A	Radiologist B	Radiologist C	Radiologist D
Elapsed time	-0.16 [-2.8; -0.05]	-0.52 [-0.59; -0.44]	-0.42 [-0.48; -0.36]	-0.23 [-0.31; -0.14]
Invalid gaze time points	0.04 [-0.07; 0.13]	-0.20 [-0.29; -0.13]	-0.20 [-0.29; -0.11]	-0.03 [-0.15; 0.16]
Rate of invalid time points	0.27 [0.18; 0.36]	-0.10 [-0.19; 0.00]	0.21 [0.12; 0.30]	-0.03 [-0.15; 0.16]
Traveled distance	-0.08 [-0.18; 0.02]	-0.54 [-0.61; -0.47]	-0.44 [-0.50; -0.38]	-0.25 [-0.33; -0.15]
Average gaze speed	0.27 [0.18; 0.36]	0.10 [0.01; 0.19]	0.13 [0.05; 0.22]	-0.08 [-0.18; 0.02]
Confidence in diagnosis	-0.01 [-0.11; 0.09]	-0.05 [-0.15; 0.05]	-0.10 [-0.21; 0.01]	0.15 [0.06; 0.25]
X-ray coverage	-0.00 [-0.11; 0.10]	-0.22 [-0.32; -0.12]	-0.53 [-0.59; -0.46]	-0.28 [-0.37; -0.20]
Lung coverage	-0.20 [-0.29; -0.10]	-0.12 [-0.21; -0.03]	-0.52 [-0.60; -0.44]	-0.28 [-0.37; -0.19]
Information gain (X-ray)	-0.04 [-0.14; 0.56]	-0.26 [-0.35; -0.16]	-0.53 [-0.59; -0.46]	-0.35 [-0.43; -0.26]
Information gain (lungs)	-0.20 [-0.30; -0.10]	-0.21 [-0.30; -0.11]	-0.52 [-0.58; -0.44]	-0.31 [-0.39; -0.22]

on the idea of consecutive analysis of airways, pleural space, cardiomediastinal contour, and bone abnormalities, which resulted in various mnemonics such as ABCDE and DRSABCDE [34]. The core idea of the guidelines is to ensure systematic coverage of all the lung regions where abnormalities can be potentially manifested. We used deep learning to automate such coverage analysis.

A deep learning algorithm for automated segmentation of lung fields from chest X-rays was implemented. We augmented a U-Net segmentation neural network [35] for contour-aware segmentation. The network was trained to map an input chest X-ray to two masks of the same size as the X-ray, where the first mask corresponded to the lung fields, and the second mask corresponded to the contour of the lung fields. By explicitly requesting both lung fields and lung contours as an output, the errors at pixels around the lung borders affected the UNet loss function more than the errors at pixels in the middle of the lung fields. The correct identification of border pixels is more contributive to the overall segmentation quality as the segmentation errors inside and outside lungs can be corrected using connected component decomposition, morphological analysis, etc. The encoder of the U-Net was replaced with a 50-layer ResNeXt [36] pre-trained on the ImageNet database. The loss function of the network was a combination of binary cross-entropy and Dice coefficient losses.

For Unet training, publicly available chest X-rays with manual segmentations were used [37], [38]. The training images were augmented using rotations of up to ±15º, shears of up to ±10º, scaling of up to ±20% to the original anatomy size, and horizontal flip as described in this work [39]. The Unet was trained with Adam optimizer with an initial learning rate of 0.001 and batch size of 16 images.

2) Gaze coverage calculation

The segmented lung fields allow us to quantitatively assess the lung coverage by the radiologist's gaze. An image coverage map is first generated from the eye-tracking data. The coordinate of the gaze for each time point $t \in T$ is defined as $(\mathbf{x}, z) \in G(t)$, where $\mathbf{x}$ is the 2D coordinate (pixels) of the gaze over the target image I, and z is the distance (mm) between the monitor and the radiologist's eyes. The visual image coverage $\psi(\mathbf{y}, t)$ at time point t, was calculated as:

$$\psi(\mathbf{y}, t) = e^{-\frac{|\mathbf{x}-\mathbf{y}|^2}{2\left(z\cdot\rho\cdot\tan\frac{\theta}{2}\right)^2}}, \quad (2)$$

where ρ is the pixel density of the monitor and $\theta = 2^{\circ}$ is the visual angle sufficient to capture abnormalities defined from the literature on medical imaging reading [40]–[42]. The gaze coverage image Ψ is calculated by accumulating $\psi(\mathbf{y}, t)$ for all image pixels:

$$\Psi(\mathbf{y}, t) = \left[\sum_{r=0}^{t} \psi(\mathbf{y}, r)\right]. \quad (3)$$

The gaze coverage image Ψ allows visualizing the area covered by the radiologist's gaze for each time point t.

3) Visual information gain histogram

We define visual information gain as a dependency between the coverage of the lung fields and time. For every time point t, the lung coverage is:

$$s(t) = \frac{\sum_{\mathbf{y}} \Psi(\mathbf{y}, t) \cdot I_{lungs}(\mathbf{y})}{\sum_{\mathbf{y}} I_{lungs}(\mathbf{y}) \cdot}, \quad (4)$$

where I_{lungs} is the binary array with the segmented lungs. The histogram of the accumulated lung coverage with the radiologist's gaze is generated for each time point t (Fig. 2). Function $s(t)$ is monotonically increasing with $s(0) = 0$, i.e., no part of the lungs is covered by the gaze at the time point zero. The intuition behind the information gain histogram is to capture two different image reading patterns: fast scanning of the image followed by focusing on the areas with potential abnormalities, and systematic image scanning. We can expect that fast image scanning from the first reading pattern will result in sharp growth of $s(t)$ followed by saturation. On the other hand, $s(t)$ will grow more linearly for systematic image scanning.

F. Statistical Analysis

The number of X-rays analyzed by a radiologist, i.e. CWD, was used as a reference metric for the fatigue analysis and

estimation of predictive powers of numerical features. The correlation between the reference CWD and image reading features, such as reading speed, eye movement speed, lung coverage, visual information gain, etc., was estimated using the Pearson correlation coefficient with $p < 0.01$ considered statistically significant. The diagnostic performance was evaluated with an F-score against the reference consensus of three radiologists who annotated the database [30]. To compute the correlation between fatigue test results and the diagnostic performance, the mean diagnostic performance was computed for the X-rays from the current database batch. Wilcoxon signed-rank test was used to compare the changes in lung coverage at the beginning and end of the experiment. The 95% confidence intervals for resulting metrics were obtained using the bootstrap method.

III. Results

A. Characteristic Features

We extracted the eye tracker data to compute various numerical features and evaluated their correlation with fatigue levels. For every X-ray reading, we calculated: 1) the elapsed time between the start and end of the reading; 2) the total gaze trajectory; 3) the average eye movement speed; 4) the lung field coverage, 5) the X-ray coverage; 6) the rate of invalid eye-tracking time points, e.g., when the eyes of the user are not looking at the screen or were not captured for other reasons; 7) blinking rate, 8) self-reported confidence in the diagnosis and 9) the mean information gain. The mean information gain was defined by normalizing the information gain histogram to the time spent reading the X-ray. The correlation coefficients between the features and the reference CWD value, i.e. the number of the X-ray in the analysis, are given in Table 2.

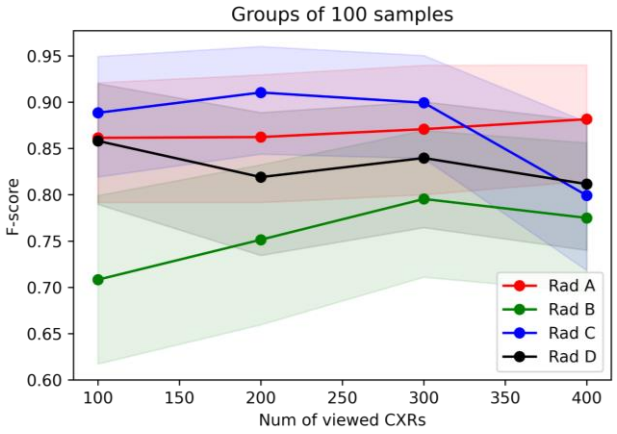


Figure 3. The performance of the radiologists in terms of diagnostic accuracy calculated for every batch of 100 X-rays. The performance of three out of four radiologists improved after the lunch break (200→300). Moreover, the performance of three out of four radiologists deteriorated at the last 100 X-rays (300→400).

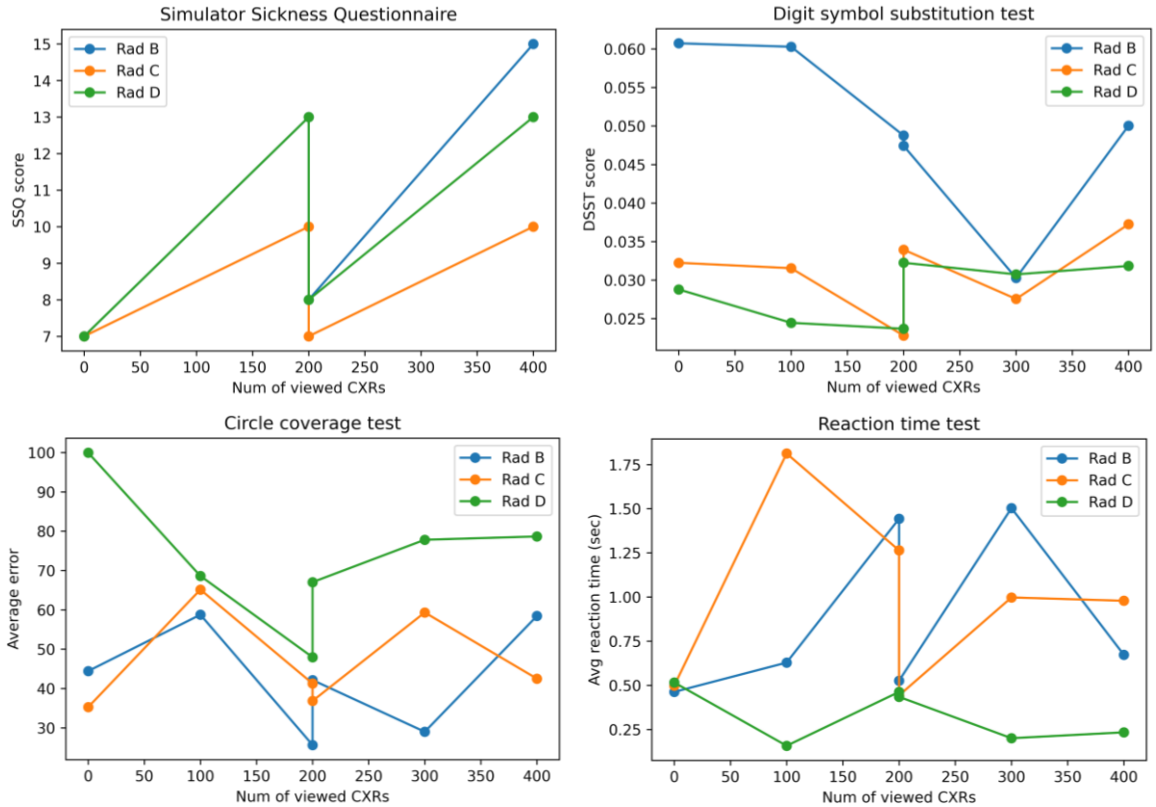


Figure 4. The results of the simulator sickness questionnaire (top-left), digit symbol substitution (top-right), circle coverage (bottom-left) and reaction time (bottom-right) tests computed for Radiologist B, C and D participating in this experiment. The tests were expected at the beginning of the experiment and after every 100 X-rays analyzed to measure the changes in fatigue, reaction and concentration. An additional series of tests was performed after a lunch break after 200 X-rays analyzed.

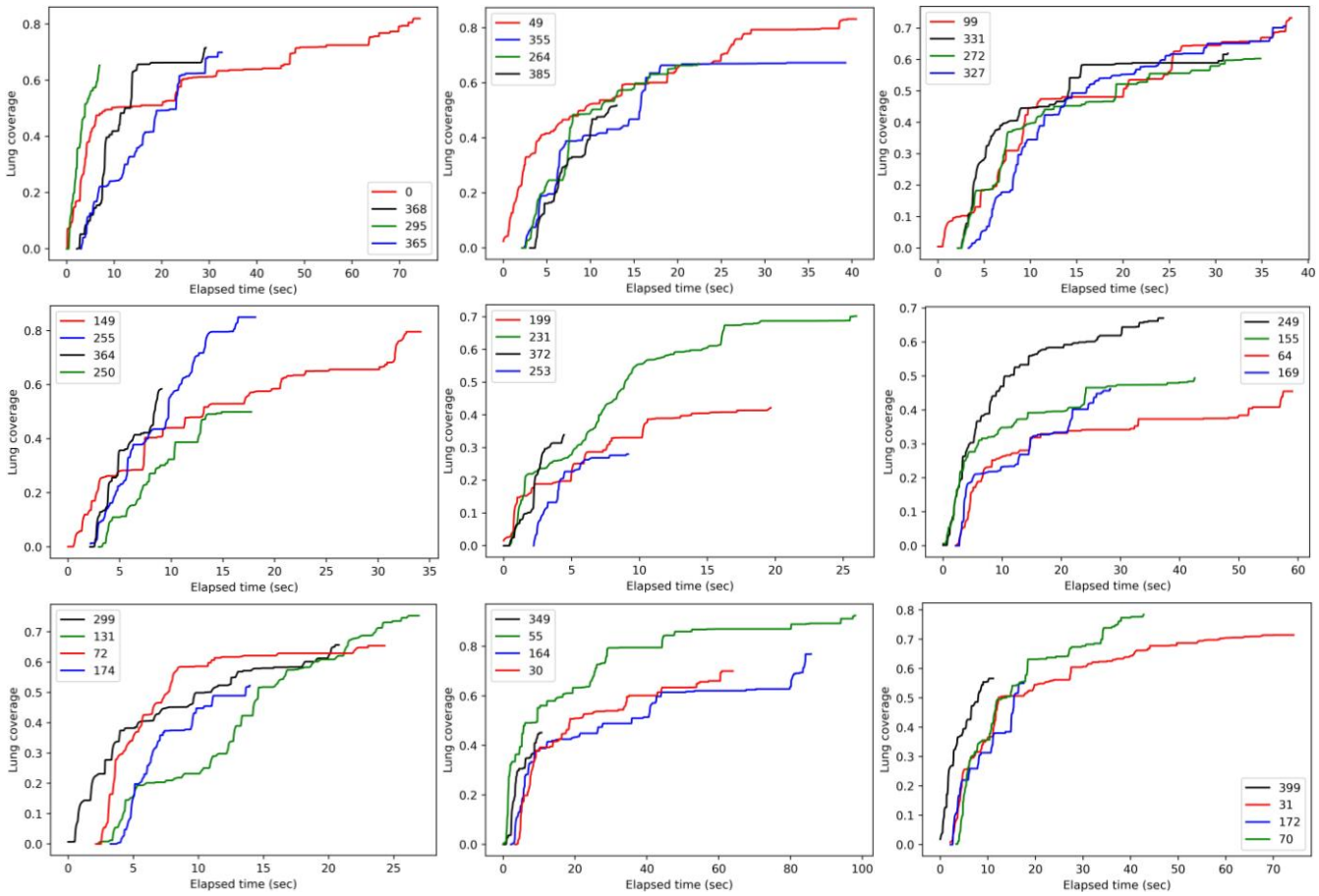


Figure 5. The comparison between the information gain histograms computed for some X-rays from the database. Each X-ray was analyzed each of the radiologist at different time point of the experiment. The order of colors in the legends of the plots correspond to the order of Radiologist A-D. The colors correspond to the order of X-rays presented: red – earliest, black - latest. For example, for left bottom plot, the corresponding X-ray was earliest presented to Radiologist C, being 72nd in his collection, second earliest presented to Radiologist B, being 131st in his collection, third earliest presented to Radiologist D, being 174th in his collection, and latest presented to Radiologist A, being 299th in his collection. The histogram comparison illustrates a visual trend of fresh radiologists (red lines) to gather more information about the depicted lungs before pronouncing the diagnosis, and of tired radiologists to be quickly satisfied after observing a small part of the lungs. For this illustration every 50th X-ray presented for Radiologist A are selected.

B. Radiologists' Diagnostic Performance

The radiologists' diagnostic performance was calculated as the average F-score for each batch, i.e., 100 chest X-rays. The diagnosis for each X-ray was manually extracted from the radiologists' voice recordings. If multiple abnormalities present, radiologists were encouraged to mention all of them during decision-making but asked to select a single abnormality as the final diagnosis. The reference diagnosis may include multiple abnormalities. The X-ray analysis was considered correct when our radiologist's final diagnosis is among the reference abnormalities or when both our radiologist and the reference team of radiologists find the X-ray to be normal. The F-scores with 95% confidence intervals calculated for each radiologist are presented in Fig 3.

The average reading time was 24.4 sec, 35.1 sec, 21.8 sec, and 25.1 sec for Radiologists A, B, C, and D, respectively. The reading times significantly correlated ($p < 0.01$) with the distance traveled by the gaze having the Person correlation coefficients of 0.93, 0.87, 0.92, and 0.74 for Radiologists A, B, C, and D, respectively. The average coverage was 64%, 65%, 58%, and 55% for lung fields and 32%, 29%, 30% and 30% for whole X-ray images for Radiologists A, B, C, and D, respectively.

C. Fatigue Measurements

Fatigue measurements computed for Radiologists B, C and D were compared to the reference CWD values and diagnostic performance. The SSQ test contains seven questions, where each question can be graded from 1 to 4 with higher grades corresponding to higher discomfort potentially associated with fatigue. Radiologists B, C, and D had the self-evaluated level of fatigue graded as 7 at the beginning of the test, i.e., no fatigue at all, and graded as 15, 8, and 13, respectively, at the end of the experiment (Fig. 4, Tab. 3). The DSST performance was evaluated as the average time needed to substitute one symbol divided by the average substitution accuracy (Fig. 4). Lower values of DSST correspond to faster and more accurate substitutions. The results of the circle coverage test were calculated as the average distance between the radiologist's gaze and the circle contour (Fig. 4). A larger distance indicates a higher deviation from the optimal trajectory. The RTT evaluated the average time needed to move to the next highlighted circle on the screen (Fig. 4).

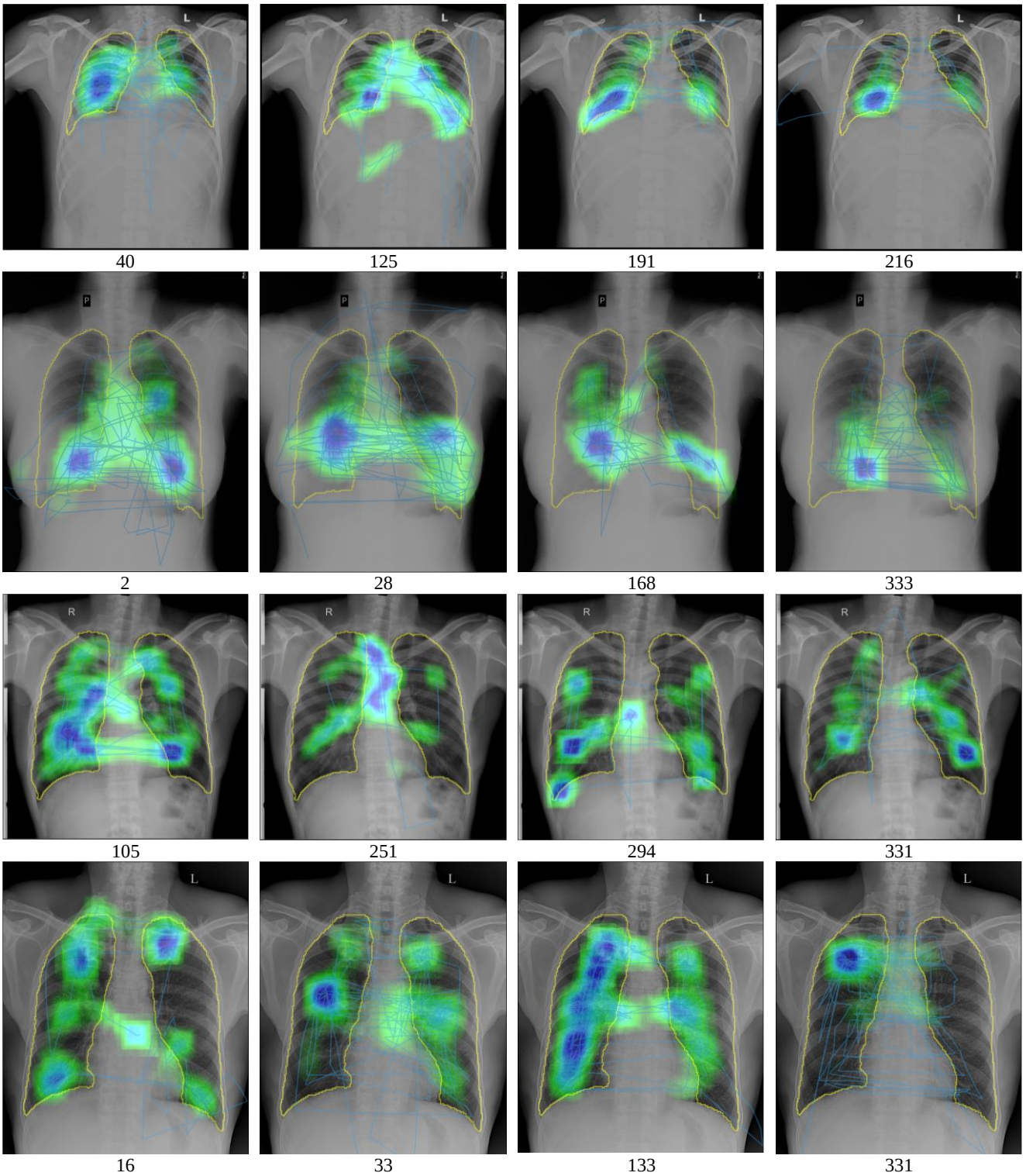


Figure 6. The comparison of gaze maps superimposed over four chest X-rays. The numbers below each X-ray indicate the order in which this X-ray was presented to a radiologist whose gaze map is depicted. Each row contains one X-ray presented to all radiologists at different time moments of the experiment.

IV. Discussion

This paper presents the first attempt at using AI for correlating radiologists' eye movements pattern and fatigue. Fatigue in radiology is believed to be one of the main sources of radiological errors [4], [23]. The challenges with fatigue analysis start with the broad definition of fatigue as an overall feeling of tiredness. Due to its inherent subjectivity, fatigue is often evaluated using self-reported questionaries such as SOFI and SSQ. Although the results of such questionaries are in general agreement with the expected fatigue growth after a day of work in radiology [23], they have significant disadvantages.

First, the questionaries are not granular enough to capture the continuity of fatigue growth. Second, the questionary results are highly subjective and do not allow reliable comparison of radiologists. The second shortcoming was manifested in our study where Radiologist C reported a minimal SSQ score growth in comparison to Radiologists B and D whose self-reported SSQ fatigue score doubled at the end of the experiment (Fig. 4). In contrast to self-reported fatigue, the diagnostic performance of Radiologist C dropped at the end of the experiment while improving for Radiologist B. It is important to note that the X-rays were given to the radiologists in a different order so Radiologist B may have got "easier" X-rays at the end.

As questionaries are highly subjective, the participants may involuntarily adjust their answers to the expected fatigue of an average person. We run three alternative tests that measure concentration and precision, which correlate with fatigue and had been used for fatigue analysis [32], [33], [43], [44]. The circle coverage test results deteriorated for Radiologists B and C, with CWD growth, using the slope of a linear regression fitted to the test performance. The results of the circle coverage test of Radiologist D also worsened with the exception of the pre-experiment measurement. The DSST and RTT exhibited inconsistent results. In contrast to the expectations, Radiologist B had improved his performance on digit substitutions in DSST, while Radiologist D - on the RTT. Such observations may indicate that the radiologists improve the concentration test performance with practice, which compensates for fatigue-related performance degradation. The potential effect of practice is supported by the fact that all three radiologists improved their performance on the second DSST attempt and two out of three radiologists improved their performance on the second RTT. We can expect that on the second attempt the radiologists have already become well acquainted with the test aims but have not yet gotten tired from X-ray reading.

Reading time exhibited a statistically significant negative correlation with the CWD for all radiologists in the experiment. The negative correlation between CWD and image reading time and positive correlations between image reading time and diagnosis quality have been continuously reported in the literature [29], [31], [45], [46]. While Krupinski et al. observed a slight reading time reduction for bone fracture diagnosis from X-ray images [31], Burling et al. found out that radiologists spend 30% less time interpreting the last five abdominal CTs than the first five during their work shifts [46]. In our study, we observed that the X-ray reading time, on average, reduces by 0.02, 0.09, 0.08, and 0.03 sec/X-ray for Radiologists A, B, C, and D, respectively. Despite a statistically significant negative correlation between reading time and the CWD for all radiologists, the reading time is not a sufficient fatigue predictor. The problem is a high inter-radiologist variability of the average reading time, which ranges from 21.8 sec for Radiologist C to 35.1 sec for Radiologist B. Consequently, the reading times for one or several X-rays for an arbitrary radiologist cannot predict the CWD without pre-collected knowledge on how fast this radiologist reads images when he/she is fresh and tired. The distance traveled by the gaze is strongly correlated to the reading time metric (correlation coefficients are 0.87-0.93) so its patterns and shortcomings are similar. The rate of invalid gaze time points is positively correlated with the reference CWD, i.e., radiologists more often distract their attention from the X-ray at the end of the experiment than at the beginning. The confidence in diagnosis insignificantly fluctuated for different radiologists (Tab. 2), but the confidence in diagnosis significantly correlated with the diagnosis quality for all radiologists

In contrast to the traveled distance, the X-ray and lung coverage is not a derivative of the elapsed time. A user could spend a lot of time looking at a small image region, which will not increase the image coverage. The lung coverage negatively correlated with the CWD for Radiologists A, C, and D, while no significant correlation was observed for Radiologist B. Using the image reading time, traveled distance, gaze speed, and lung coverage statistics (Tab. 2), we can reconstruct radiologists' image reading pattern when tired. Radiologist A starts to spend less time but tries to rapidly cover the images with his gaze. While continuing to cover the X-rays on the same level, Radiologist A covers the lungs significantly less with CWD growth. This observation suggests that his image reading becomes less focused with time. The gaze speed for Radiologists B, C, and D mildly change during the experiment, which, considering the shortening reading time, resulted in a reduction of X-ray coverage with CWD growth. Radiologist C demonstrated the most considerable drop of the X-ray and lung coverage which is accompanied by the highest relative drop in performance (Fig 4) among all radiologists.

The information gain turned out to be the most reliable predictor for the CWD in the experiment. The metric calculated for lung fields exhibited a statistically significant correlation with fatigue for all radiologists. The correlation coefficients suggest that the radiologists need less information to make a decision at the end of the experiment than at the beginning. To confirm this explanation, we plotted the information gain curves for all radiologists for several X-rays (Fig 5). As the X-rays are shuffled for each radiologist, the same X-ray could be analyzed closer to the beginning or end by different radiologists. There is a clear pattern that the radiologists try to gain more information about the X-ray at the beginning of the experiment than at the end (Fig. 6). The average information gain of 0.455 for the first batch of X-rays is statistically significantly higher than the average information of 0.332 for the last batch.

One of the key advantages of the information gain is that this metric blends the reading time, speed, and object coverage feature into a single value. Such a metric can mitigate the individual X-ray reading patterns observed for tired radiologists: faster but imprecise gaze movements (Radiologist A), faster reading with low (Radiologist B) or high (Radiologist C) number of invalid gaze time points, and faster reading with slower gaze movements (Radiologist D).

V. Conclusion

We have demonstrated that deep-learning-based segmentation of anatomical structures on medical images, in our case segmentation of lung fields from chest X-rays, can be extended further than classical computer-aided diagnosis applications. We used lung field segmentation to quantitatively assess the lung coverage by radiologists' gaze and mapped this with radiologists' CWD. Such distilling of the gaze over the

complete image to the gaze over target areas can become essential for further radiologists' performance evaluation. Despite correlations with the CDW values for some concentration tests, they did not exhibit significant fatigue-predictive powers. This fact does not mean that the tests are not applicable, but a more elaborated analysis is needed to exclude potential factors of improved performance due to practice. The presented study has several limitations associated with the controlled setting of the experiment. In clinical practice, radiologists have irregular breaks during image reading, spend time on clinical notes, consult with colleagues, etc. Our settings are therefore not truly clinical. At the same time, a standardized setup was necessary for a reliable comparison between the experiment participants. Another limitation is that we selected abnormalities that we unanimously diagnosed by a team of radiologists who generated the public database of chest X-rays. Although such a unanimity suggests the high quality and reliability of the reference labels, it could also indicate that the cases might be easier and faster to diagnose than cases with unclear diagnoses. The differences in radiologists' behavior for easy and challenging cases will be analyzed in future work. The future work will also be focused on a more detailed analysis of the eye movement features, such as fixation duration and saccade properties [47], and their correlation with radiologists' performance.

REFERENCES

[1] R. J. M. Bruls and R. M. Kwee, "Workload for radiologists during on-call hours: dramatic increase in the past 15 years," *Insights Imaging*, vol. 11, no. 1, p. 121, Nov. 2020, doi: 10.1186/s13244-020-00925-z.

[2] S. K. Mun, K. H. Wong, S.-C. B. Lo, Y. Li, and S. Bayarsaikhan, "Artificial Intelligence for the Future Radiology Diagnostic Service," *Front. Mol. Biosci.*, vol. 7, p. 512, 2021, doi: 10.3389/fmolb.2020.614258.

[3] J. Born *et al.*, "On the role of artificial intelligence in medical imaging of COVID-19," *Patterns N. Y. N*, vol. 2, no. 6, p. 100269, Jun. 2021, doi: 10.1016/j.patter.2021.100269.

[4] M. A. Bruno, E. A. Walker, and H. H. Abujudeh, "Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction," *Radiogr. Rev. Publ. Radiol. Soc. N. Am. Inc*, vol. 35, no. 6, pp. 1668–1676, Oct. 2015, doi: 10.1148/rg.2015150023.

[5] L. Berlin, "Radiologic errors, past, present and future," *Diagn. Berl. Ger.*, vol. 1, no. 1, pp. 79–84, Jan. 2014, doi: 10.1515/dx-2013-0012.

[6] S. Littlefair, P. Brennan, W. Reed, and C. Mello-Thoms, "Does Expectation of Abnormality Affect the Search Pattern of Radiologists When Looking for Pulmonary Nodules?," *J. Digit. Imaging*, vol. 30, no. 1, pp. 55–62, Feb. 2017, doi: 10.1007/s10278-016-9908-7.

[7] J. Chen, S. Littlefair, R. Bourne, and W. M. Reed, "The Effect of Visual Hindsight Bias on Radiologist Perception," *Acad. Radiol.*, vol. 27, no. 7, pp. 977–984, Jul. 2020, doi: 10.1016/j.acra.2019.09.032.

[8] S. Waite, J. Scott, B. Gale, T. Fuchs, S. Kolla, and D. Reede, "Interpretive Error in Radiology," *AJR Am. J. Roentgenol.*, vol. 208, no. 4, pp. 739–749, Apr. 2017, doi: 10.2214/AJR.16.16963.

[9] N. Stec, D. Arje, A. R. Moody, E. A. Krupinski, and P. N. Tyrrell, "A Systematic Review of Fatigue in Radiology: Is It a Problem?," *Am. J. Roentgenol.*, vol. 210, no. 4, pp. 799–806, Apr. 2018, doi: 10.2214/AJR.17.18613.

[10] S. Taylor-Phillips and C. Stinton, "Fatigue in radiology: a fertile area for future research," *Br. J. Radiol.*, vol. 92, no. 1099, p. 20190043, Jul. 2019, doi: 10.1259/bjr.20190043.

[11] T. N. Hanna *et al.*, "The Effects of Fatigue From Overnight Shifts on Radiology Search Patterns and Diagnostic Performance," *J. Am. Coll. Radiol. JACR*, vol. 15, no. 12, pp. 1709–1716, Dec. 2018, doi: 10.1016/j.jacr.2017.12.019.

[12] Y. Ikushima, H. Yabuuchi, H. Honda, and J. Morishita, "SU-E-I-62: Investigation of Dominant Factors Affecting Fatigue in Image Reading of Radiologists," *Med. Phys.*, vol. 39, no. 6Part5, pp. 3639–3639, 2012, doi: 10.1118/1.4734778.

[13] T. Nihashi *et al.*, "Monitoring of fatigue in radiologists during prolonged image interpretation using fNIRS," *Jpn. J. Radiol.*, vol. 37, no. 6, pp. 437–448, Jun. 2019, doi: 10.1007/s11604-019-00826-2.

[14] N. Z. Ndaro and S.-Y. Wang, "Effects of Fatigue Based on Electroencephalography Signal during Laparoscopic Surgical Simulation," *Minim. Invasive Surg.*, vol. 2018, p. e2389158, May 2018, doi: 10.1155/2018/2389158.

[15] L. Lévêque, H. Bosmans, L. Cockmartin, and H. Liu, "State of the Art: Eye-Tracking Studies in Medical Imaging," *IEEE Access*, vol. 6, pp. 37023–37034, 2018, doi: 10.1109/ACCESS.2018.2851451.

[16] C. E. Connor, H. E. Egeth, and S. Yantis, "Visual Attention: Bottom-Up Versus Top-Down," *Curr. Biol.*, vol. 14, no. 19, pp. R850–R852, Oct. 2004, doi: 10.1016/j.cub.2004.09.041.

[17] H. L. Kundel, C. F. Nodine, E. A. Krupinski, and C. Mello-Thoms, "Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms," *Acad. Radiol.*, vol. 15, no. 7, pp. 881–886, Jul. 2008, doi: 10.1016/j.acra.2008.01.023.

[18] D. P. Turgeon and E. W. N. Lam, "Influence of Experience and Training on Dental Students' Examination Performance Regarding Panoramic Images," *J. Dent. Educ.*, vol. 80, no. 2, pp. 156–164, Feb. 2016.

[19] B. Law, M. S. Atkins, A. E. Kirkpatrick, and A. J. Lomax, "Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment," in *Proceedings of the 2004 symposium on Eye tracking research & applications*, New York, NY, USA, Mar. 2004, pp. 41–48. doi: 10.1145/968363.968370.

[20] L. Cooper, A. Gale, I. Darker, A. Toms, and J. Saada, "Radiology image perception and observer performance: How does expertise and clinical information alter interpretation? Stroke detection explored through eye-tracking," vol. 7263, p. 72630K, Feb. 2009, doi: 10.1117/12.811098.

[21] K. Saab *et al.*, "Observational Supervision for Medical Image Classification Using Gaze Data," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Cham, 2021, pp. 603–614. doi: 10.1007/978-3-030-87196-3_56.

[22] A. Karargyris *et al.*, "Creation and validation of a chest X-ray dataset with eye-tracking and report dictation for AI development," *Sci. Data*, vol. 8, no. 1, Art. no. 1, Mar. 2021, doi: 10.1038/s41597-021-00863-5.

[23] M. Bhattacharya, S. Jain, and P. Prasanna, "RadioTransformer: A Cascaded Global-Focal Transformer for Visual Attention-guided Disease Classification," *ArXiv220211781 Cs*, Feb. 2022, Accessed: May 05, 2022. [Online]. Available: http://arxiv.org/abs/2202.11781

[24] S. Waite *et al.*, "Tired in the Reading Room: The Influence of Fatigue in Radiology," *J. Am. Coll. Radiol. JACR*, vol. 14, no. 2, pp. 191–197, Feb. 2017, doi: 10.1016/j.jacr.2016.10.009.

[25] R. J. Leigh and D. S. Zee, *The Neurology of Eye Movements*. Oxford University Press. Accessed: Nov. 23, 2021. [Online]. Available: https://oxfordmedicine.com/view/10.1093/med/9780199969289.001.0001/med-9780199969289;jsessionid=F751D2A22711ADF43B5CA3ED6D008E07

[26] M. K. Eckstein, B. Guerra-Carrillo, A. T. Miller Singley, and S. A. Bunge, "Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development?," *Dev. Cogn. Neurosci.*, vol. 25, pp. 69–91, Jun. 2017, doi: 10.1016/j.dcn.2016.11.001.

[27] B. Zheng, X. Jiang, and M. S. Atkins, "Detection of Changes in Surgical Difficulty: Evidence From Pupil Responses," *Surg. Innov.*, vol. 22, no. 6, pp. 629–635, Dec. 2015, doi: 10.1177/1553350615573582.

[28] T. T. Brunyé, M. D. Eddy, E. Mercan, K. H. Allison, D. L. Weaver, and J. G. Elmore, "Pupil diameter changes reflect difficulty and diagnostic accuracy during medical image interpretation," *BMC Med. Inform. Decis. Mak.*, vol. 16, p. 77, Jul. 2016, doi: 10.1186/s12911-016-0322-3.

[29] A. van der Gijp *et al.*, "How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology," *Adv. Health Sci. Educ. Theory Pract.*, vol. 22, no. 3, pp. 765–787, Aug. 2017, doi: 10.1007/s10459-016-9698-1.

[30] H. Q. Nguyen *et al.*, "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations," *ArXiv201215029 Eess*, Jan. 2021, Accessed: Nov. 05, 2021. [Online]. Available: http://arxiv.org/abs/2012.15029

[31] E. A. Krupinski, K. S. Berbaum, R. T. Caldwell, K. M. Schartz, and J. Kim, "Long Radiology Workdays Reduce Detection and

Accommodation Accuracy," *J. Am. Coll. Radiol.*, vol. 7, no. 9, pp. 698–704, Sep. 2010, doi: 10.1016/j.jacr.2010.03.004.

[32] J. Jaeger, "Digit Symbol Substitution Test," *J. Clin. Psychopharmacol.*, vol. 38, no. 5, pp. 513–519, Oct. 2018, doi: 10.1097/JCP.0000000000000941.

[33] V. W.-S. Tseng, N. Valliappan, V. Ramachandran, T. Choudhury, and V. Navalpakkam, "Digital biomarker of mental fatigue," *Npj Digit. Med.*, vol. 4, no. 1, pp. 1–5, Mar. 2021, doi: 10.1038/s41746-021-00415-6.

[34] J. S. Klein and M. L. Rosado-de-Christenson, "A Systematic Approach to Chest Radiographic Analysis," in *Diseases of the Chest, Breast, Heart and Vessels 2019-2022: Diagnostic and Interventional Imaging*, J. Hodler, R. A. Kubik-Huch, and G. K. von Schulthess, Eds. Cham (CH): Springer, 2019. Accessed: Nov. 08, 2021. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK553874/

[35] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Cham, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[36] S. Xie *et al.*, "Artifact Removal using Improved GoogLeNet for Sparse-view CT Reconstruction," *Sci. Rep.*, vol. 8, no. 1, p. 6700, Dec. 2018, doi: 10.1038/s41598-018-25153-w.

[37] B. van Ginneken, M. B. Stegmann, and M. Loog, "Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database," *Med. Image Anal.*, vol. 10, no. 1, pp. 19–40, Feb. 2006, doi: 10.1016/j.media.2005.02.002.

[38] J. Shiraishi *et al.*, "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," *AJR Am. J. Roentgenol.*, vol. 174, no. 1, pp. 71–74, Jan. 2000, doi: 10.2214/ajr.174.1.1740071.

[39] I. Sirazitdinov, M. Kholiavchenko, R. Kuleev, and B. Ibragimov, "Data Augmentation for Chest Pathologies Classification," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, Apr. 2019, pp. 1216–1219. doi: 10.1109/ISBI.2019.8759573.

[40] H. L. Kundel, C. F. Nodine, D. Thickman, and L. Toto, "Searching for lung nodules. A comparison of human performance with random and systematic scanning models," *Invest. Radiol.*, vol. 22, no. 5, pp. 417–422, May 1987, doi: 10.1097/00004424-198705000-00010.

[41] J. M. Wolfe, C.-C. Wu, J. Li, and S. B. Suresh, "What do experts look at and what do experts find when reading mammograms?," *J. Med. Imaging Bellingham Wash*, vol. 8, no. 4, p. 045501, Jul. 2021, doi: 10.1117/1.JMI.8.4.045501.

[42] H. Strasburger, I. Rentschler, and M. Jüttner, "Peripheral vision and pattern recognition: A review," *J. Vis.*, vol. 11, no. 5, p. 13, Dec. 2011, doi: 10.1167/11.5.13.

[43] Z. Guo, R. Chen, K. Zhang, Y. Pan, and J. Wu, "The Impairing Effect of Mental Fatigue on Visual Sustained Attention under Monotonous Multi-Object Visual Attention Task in Long Durations: An Event-Related Potential Based Study," *PLOS ONE*, vol. 11, no. 9, p. e0163360, Sep. 2016, doi: 10.1371/journal.pone.0163360.

[44] L. G. Faber, N. M. Maurits, and M. M. Lorist, "Mental Fatigue Affects Visual Selective Attention," *PLOS ONE*, vol. 7, no. 10, p. e48073, Oct. 2012, doi: 10.1371/journal.pone.0048073.

[45] E. Sokolovskaya *et al.*, "The Effect of Faster Reporting Speed for Imaging Studies on the Number of Misses and Interpretation Errors: A Pilot Study," *J. Am. Coll. Radiol. JACR*, vol. 12, no. 7, pp. 683–688, Jul. 2015, doi: 10.1016/j.jacr.2015.03.040.

[46] D. Burling *et al.*, "CT colonography interpretation times: effect of reader experience, fatigue, and scan findings in a multi-centre setting," *Eur. Radiol.*, vol. 16, no. 8, pp. 1745–1749, Aug. 2006, doi: 10.1007/s00330-006-0190-9.

[47] J. Lou, X. Zhao, P. Young, R. White, and H. Liu, "Study of Saccadic Eye Movements in Diagnostic Imaging," in *2021 IEEE International Conference on Image Processing (ICIP)*, Sep. 2021, pp. 1474–1478. doi: 10.1109/ICIP42928.2021.9506017.