AI-based analysis of radiologist's eye movements for fatigue estimation: A pilot study on chest X-rays

Ilya Pershin[a], Maksim Kholiavchenko[b], Bulat Maksudov[a], Tamerlan Mustafaev[a], and Bulat Ibragimov[a,d]

[a]Innopolis University, Innopolis city, Russia
[b]Rensselaer Polytechnic Institute, New York, USA
[d]University of Copenhagen, Copenhagen, Denmark

ABSTRACT

Radiologist-AI interaction is a novel area of research of potentially great impact. It has been observed in the literature that the radiologists' performance deteriorates towards the shift ends and there is a visual change in their gaze patterns. However, the quantitative features in these patterns that would be predictive of fatigue have not yet been discovered. A radiologist was recruited to read chest X-rays, while his eye movements were recorded. His fatigue was measured using the target concentration test and Stroop test having the number of analyzed X-rays being the reference fatigue metric. A framework with two convolutional neural networks based on UNet and ResNeXt50 architectures was developed for the segmentation of lung fields. This segmentation was used to analyze radiologist's gaze patterns. With a correlation coefficient of 0.82, the eye gaze features extracted lung segmentation exhibited the strongest fatigue predictive powers in contrast to alternative features.

Keywords: eye tracking, deep learning, lung fields, chest

1. INTRODUCTION

Medical imaging is the main tool in the work of the radiologist. In recent years, there has been no significant reduction in errors in radiological studies.[1] In addition, workload increases without a growth in the number of staff, which impacts a decrease in the quality of diagnostics.[2] Modern advances in artificial intelligence technologies make it easier for the radiologist to work, which was especially clearly demonstrated during the COVID-19 pandemic.[3] Nevertheless, today it is necessary to better understand the process of visual perception by a radiologist for to improve modern computer-aided detection systems.[4]

The quality of radiation diagnostics depends on many factors, including fatigue.[5] Modern methods of measuring fatigue are based on electrocardiograms, electromyograms, galvanic skin reactions and electroencephalography,[6] but these methods require mounting sensors on the participant's body. The main advantage of using eye tracking is the study of neurophysiology and human psychology without attaching special sensors to the body. In this paper we present approach for estimation fatigue of radiologist based on eye-tracking data and deep learning methods. We assume that the reference fatigue level decreases linearly over time, and comparison the proposed approach with Stroop and target concentration test.

2. METHODOLOGY

2.1 Experiment setting

Chest X-rays from three public databases were used in this study. In particular, 400 X-rays were randomly sampled from the databases CheXpert,[7] RSNA,[8] and SIIM-ACR[9] databases. From these databases, 60 X-rays with pneumonia, 60 X-rays with pneumothorax, 120 X-rays with various other abnormalities, and 160 X-rays with no abnormality labels were extracted.

Further author information: (Send correspondence to Bulat Ibragimov)
Bulat Ibragimov: E-mail: bulat@di.ku.dk

An in-house framework mimicking a radiologist workstation was installed at our research facility. A practicing radiologist with experience in X-ray analysis was recruited to participate in the gaze analysis experiment. The radiologist was unaware of the experiment's aims except that we wanted to record his eye movements during work. The framework was equipped with a diagnostic graphical user interface (GUI) and performed eye movement recording using Tobii Eye Tracker 4C and voice recording to collect radiologist activity data. The radiologist could use the mouse controller to adjust image contrast during reading, which was also recorded by the framework. The radiologist used GUI to select the appropriate diagnosis and his confidence level. Fig.1 (left) shows an example of a gaze path and gaze heatmap over the chest X-ray.

2.2 Experiment execution

The radiologist did not have a night shift before the experiment to ensure that he was fresh at the experiment start and his fatigue built up mainly due to X-ray reading. The X-rays were analyzed in batches of 100 X-rays separated by fatigue measurements and short breaks with documented duration. The level of fatigue was also measured at the start and end of the experiment. After the analysis of 200 X-rays, the radiologist had a lunch break. Three types of fatigue evaluation were documented.

The first evaluation was performed using the target concentration test.[10] The radiologist looked at a target – a set of colored concentric circles – and tried to focus his gaze on the most inner circle for a predefined time. The inner, middle and outer circles were of radius 5 mm, 17.5 mm, and 35 mm, respectively. The eye-tracking equipment measured the focusing concentration. The second evaluation was performed using the Stroop color-word test. The radiologist was iteratively shown 40 samples of colored words 'red', 'green', 'blue' and 'yellow'. The word coloring could be congruent or incongruent. For each word, the radiologist was asked to select its color. The time needed to give a correct answer to the word coloring was recorded. The reference fatigue evaluation was derived from the cumulative number of the analyzed X-rays, i.e., the first and last X-rays are assigned the lowest and highest fatigue value, respectively.

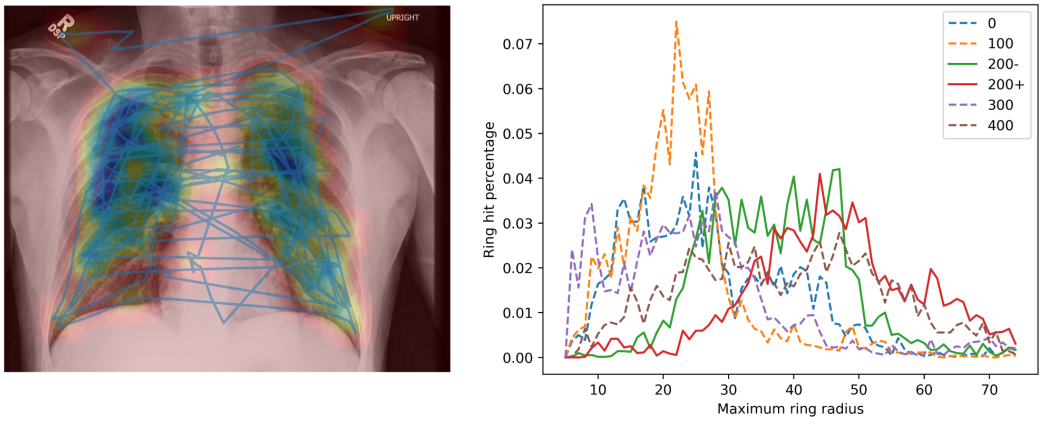

Figure 1. (left) Radiologist's gaze heatmap and gaze path superimposed over the chest X-ray. (right) Histogram with the cumulative percentage of the gaze hitting areas located in 5-40mm proximity of the target center.

2.3 Deep learning framework for radiologist's gaze assessment

Despite the chest X-ray reading style is very personalized, international recommendations have been developed to ensure a comprehensive analysis of each X-ray, e.g. the ABCDE approach.[11] The idea of our framework is to automatically recognize if all the lung field areas have been sufficiently viewed by the radiologist and to investigate if fatigue is manifested in abnormalities in the X-ray reading patterns.

We implemented a contour-aware lung segmentation algorithm,[12] and trained it on a publicly available database.[13] The database consists of 247 chest X-rays with both rights and left lungs manually segmented. A convolutional neural network (CNN) was designed to map a 1-channel input 2D chest X-ray with 2-channel

output where the channels represent the lung segmentation mask and lung segmentation borders, respectively. The UNet[14] segmentation architecture was used for the CNN.

3. EXPERIMENTS AND RESULTS

3.1 Numerical features for performance evaluation

The predictive powers of the proposed framework were evaluated in contrast to various numerical features that characterize radiologists' reading. Two time features were calculated, namely a) the time used to analyze an X-ray, and b) the time spent on X-ray contrast adjustment. Four radiologist inputs were used as features, namely a) the assigned presence of abnormality; b) the assigned presence of multiple pathologies; d) confidence in the diagnosis. Five gaze features were used, namely a) blink rate; b) the gaze heatmap area, i.e., how large was the X-ray part that received the radiologist's attention; c) the total distance traveled by the radiologist's gaze; d) the average distance from the radiologist's eyes and monitor e) the proportion of invalid eye-tracking timestamps, e.g., when the eyes were not looking at the screen.

The correlation coefficient between the reference fatigue levels and each numerical feature and features from the anatomical descriptor was computed. The coefficients were computed for all X-ray, only healthy X-ray, and only pathological X-ray. Finally, we aggregated features for several consecutive image readings to compensate for local fluctuations of gaze patterns. The obtained results are presented in Fig. 2.

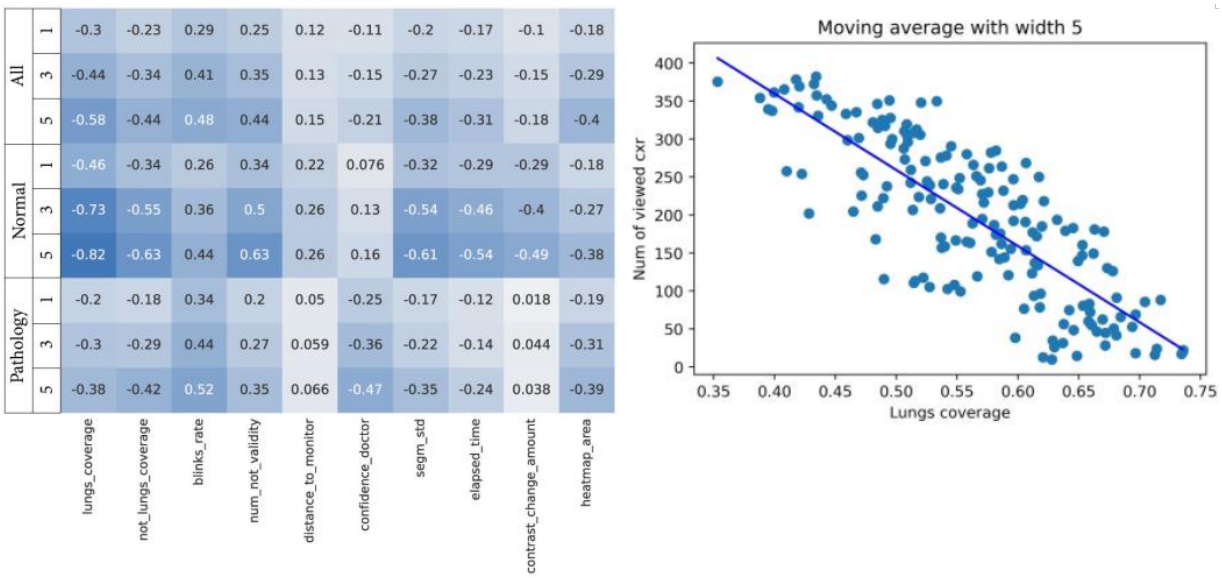

Figure 2. (left) The correlation coefficients between the reference label of radiologist fatigue and features computed for fatigue estimation. The first two features correspond to the radiologist's gaze coverage of the lungs and outside area computed using the developed deep learning framework. (right) The relationships between cumulative number of X-ray viewed and lung coverage with radiologist's gaze.

3.2 Comparison of fatigue metrics

After the target concentration test, the radiologist's gaze was converted into a gaze heatmap with high intensities at the areas that received the most attention. The following numerical features were calculated: distance between the center of the heatmap and the target, standard deviation of heatmap and gaze coordinates. We computed the histogram with the cumulative percentage of the gaze hitting areas located in 5-40mm proximity of the target center (see Fig.1 (right)). Such a histogram is similar to a dose-volume histogram for radiotherapy planning. The volume of the histogram was used as a metric. From the Stroop test, we first computed the mean time needed for the correct answer for congruent and incongruent words. The difference between the mean times –

called the Stroop effect – was used as a metric. The metrics from the target concentration and Stroop tests were compared to the reference fatigue metric. The corresponding correlation coefficients were -0.53, 0.35, 0.39, 0.6, and 0.13 for the distance between the target circle and heat map mass center, standard deviation of the heat map, the standard deviation of the gaze, cumulative histograms volumes, and the Stroop effect.

4. CONCLUSION

In this study, we presented a pilot work on the use of deep learning for the analysis of radiologist's gaze coverage and fatigue prediction. We conclude that the target concentration test could be used to measure fatigue level, while the Stroop test does not exhibit predictive powers. The most critical conclusion goes to the fact that the features generated from the proposed deep learning-based framework exhibited the strongest correlation with the fatigue levels. From Fig. 2 (left) it is clear that the lung coverage estimated from the segmented lungs cannot be replaced by simpler metrics such as the elapsed time, the gaze traveled distance, and the gaze heatmap area. We hypothesize that a tired radiologist's gaze becomes less focused on the objects of interest and travels over other image parts.

We acknowledge that the use of a single radiologist in this pilot study is a limitation. This limitation is partially compensated by the use of a very robust machine learning methodology for predictive feature selection. The inclusion of additional radiologists and more X-rays is the direction for our future research.

ACKNOWLEDGMENTS

This research was supported by the Russian Science Foundation under Grant no.18-71-10072.

REFERENCES

[1] Waite, S., Scott, J., Fuchs, T., Kolla, S., and Reede, D., "Interpretive error in radiology," *American Journal of Roentgenology* **208**, 739–749 (2017).
[2] Bruls, R. and Kwee, R., "Workload for radiologists during on-call hours: dramatic increase in the past 15 years," *Insights Imaging 11* **121** (2020).
[3] Bien, S. T. et al., "Rsna international trends: A global perspective on the covid-19 pandemic and radiology in late 2020," *Radiology* **299** (2020).
[4] Waite, S. et al., "Analysis of perceptual expertise in radiology – current knowledge and a new perspective," *Frontiers in Human Neuroscience* **13**, 213 (2019).
[5] Taylor-Phillips, S. and Stinton, C., "Fatigue in radiology: a fertile area for future research," *The British Journal of Radiology* **92**(1099), 20190043 (2019).
[6] Zeng, Z. et al., "Nonintrusive monitoring of mental fatigue status using epidermal electronic systems and machine-learning algorithms," *ACS Sensors* **5**(5), 1305–1313 (2020).
[7] Irvin, J. et al., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in [*Proceedings of the AAAI Conference on Artificial Intelligence*], **33**, 590–597 (2019).
[8] Wang, X. et al., "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , 3462–3471 (2017).
[9] Filice, R. et al., "Crowdsourcing pneumothorax annotations using machine learning annotations on the nih chest x-ray dataset," *J Digit Imaging.* **33**, 490–496 (2020).
[10] Chang, K.-M. and Chueh, M.-T. W., "Using eye tracking to assess gaze concentration in meditation," *Sensors* **19**(7) (2019).
[11] Sheard, S., [*The Chest X-ray: a Survival Guide*] (2009).
[12] Kholiavchenko, M. et al., "Contour-aware multi-label chest x-ray organ segmentation," *Int J Comput Assist Radiol Surg.* **15**, 425–436 (2020).
[13] van Ginneken, B., Stegmann, M., and Loog, M., "Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database," *Med Image Anal.* **10**, 19–40 (2006).
[14] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," *CoRR* (2015).