2024 IEEE International Symposium on Biomedical Imaging (ISBI) | 979-8-3503-1333-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISBI56570.2024.10635445

LOCALISATION OF RACIAL INFORMATION IN CHEST X-RAY FOR DEEP LEARNING DIAGNOSIS

Olivier Salvado[1,2], *Salamata Konate*[1,2], *Rodrigo Santa Cruz*[1,2], *Andrew Bdadley*[2]
Judy Wawira Gichoya[3], *Laleh Seyyed-Kalantari*[4], *Brandon Price*[5], *Clinton Fookes*[2], *Léo Lebrat*[1,2]

[1] Imaging and Computer Vision Group, CSIRO Data61, Australia
[2] SAIVT, Queensland University of Technology, Australia
[3] Department of Radiology and Imaging Sciences, Emory University, USA
[4] Electrical Engineering and Computer Science, York University, CA
[5] College of Medicine, Florida State University, USA
olivier.salvado@data61.csiro.au

ABSTRACT

Deep learning-based classification of diseases from Chest X-ray has been shown to use implicit information about the subjects' self-reported race, which could result in diagnostic bias. In this paper, we describe and compare two approaches to investigate where racial information is located in the image: first leveraging non-linear registration and computing atlas differences and second using saliency maps. We compute a map visualising the racial information between black and white subjects and discuss whether those maps are consistent with the model explanation.

Index Terms— saliency maps, eXplainable AI (XAI), chest, x-ray, deep learning, atlas-based registration.

1. INTRODUCTION

Chest X-ray is a common and relatively inexpensive medical exam for investigating many disorders and is well suited for automated analysis using deep learning [1]. However, a recent study indicates that a deep learning model trained to classify diseases also learns information about the self-reported race of the subject [2]. Naturally, this raises concerns due to the potential consequences of racial bias during diagnosis.

Recent investigations did not reveal where the model extracted racial information in the image, something that human experts could not visually reproduce [2], despite investigating all available covariates and many image features. In this paper, we expand on those investigations, and try to identify where in the chest X-ray racial information is located using explainable methods.

We use the Gichoya et al. [2] trained deep learning model that classifies the subjects' race with an accuracy of 97%. We then compared 2 approaches. First, by building a race-specific registration atlas (for white and black subjects) and comparing them using bootstrap statistics. Second, we compare averaged saliency maps generated by four recent interpretability methods.

We show that both approaches provide consistent results and point towards a similar area of the image that could hold racial information about the subjects. We first explain our data and methods before describing the experiments that we conducted. Finally, we discuss the main results.

2. METHODS AND EXPERIMENTS

2.1. Data

The data from the RSNA-CXR dataset [3] was used for comparing saliency methods. An EfficientNet Model was trained [4] to classify no-finding from lung opacity, using 500 images in each class. For the race study, the data was sourced from the MIMIC-CXR open-source chest X-ray dataset [5]. It includes information about race and other relevant clinical information. We randomly selected 1000 scans from black subjects (BS) and 1000 scans from white subjects (WS) for testing. Racial information was self-reported by subjects. All images that included foreign objects (wires, tags, implants, etc.) were manually excluded from this study.

2.2. Saliency maps

Several saliency methods were employed using the TorchRay package [6], and we selected 3 with the best qualitative results after experimentation: Gradcam [7], Extremal Perturbation [6], and Rise [8]. The perturbation method required selecting the proportion of the image to be perturbed, and we used 10%.

We also used the RankPix method, which was implemented with a slight modification [9].

RankPix requires to arbitrarily define the layers of a model where multiplicative binary masks are computed. Each

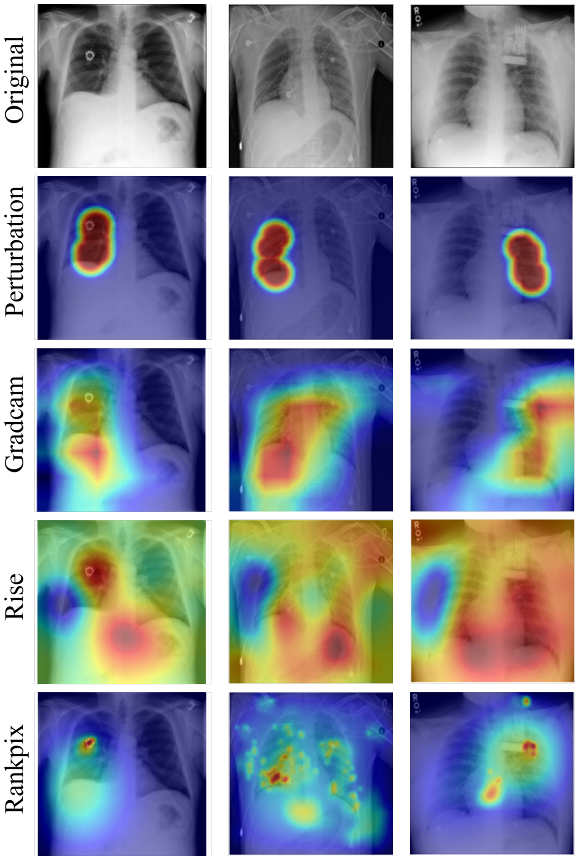

Fig. 1. Comparison of 4 saliency methods (lines) for 3 chest X-ray cases (columns) for healthy subjects with negative diagnosis but that included foreign objects (wire, devices,...). Most methods highlight the expected location, with GradCam and RankPix providing superior results.

mask defines the layer features that are passed through to the next layer. In our experiments, we used 8 layers distributed along the depth of the model.

Starting with the deepest layer (closer to the classifier head), the mask is initialised with "1", meaning all the pixels of that layer are used to compute the model output. Then, each pixel is visited and changed to 0 while calculating the change in the model output. The pixel with the most difference is selected and kept at 0, interpreting it as the most important pixel to explain the model's prediction. Then, the second most important pixel is selected using a greedy approach on the remaining "1" pixels. This process is repeated for that layer until the model output changes its prediction (i.e. from Black to White or *vice versa*).

Once a mask for a layer is computed, it is interpolated to the previous layer using transpose convolution taking into account the kernel size of the layers (all 1s). A similar process of visiting pixels and retaining the ones changing the model output is repeated for the new layer, except that the pixels being visited are only the ones within the interpolated mask of the previous layer.

Finally, all the masks are up-sampled to the full image resolution (still using transpose convolution with the corresponding kernel size) and averaged together to build the final saliency map, showing the most important area of the image that changes the model's prediction (explaining the model output). More details can be found in the original paper [9].

2.3. Comparison of saliency methods

Using the RSNA-CXR dataset, training of a ResNet34 reached an accuracy of 97.99% to classify images with no-finding from lung opacity. For training, the no-finding class excluded any images with visible objects (wires, tags, devices, ...). We then tested images with no finding where inorganic objects were present to investigate whether the saliency maps would show those objects not included in the training. Typical qualitative results are shown in Fig. 1. The RankPix method showed sharper, smaller, and more accurate localisation of the object(s) assessed visually by the authors.

2.4. Population atlases

We then investigated race localisation using the MIMIC-CXR dataset. Our goal was to compare the averages of saliency maps between WS and BS, but for that a resampling of all the saliency maps to a common template was needed. We thus built a population template.

First, all images were registered to a randomly selected image using affine registration and averaged to build an initial population template. Then all images were registered again to this template using non-linear registration, and averaged to construct a new population template. This process was repeated 5 times to yield the final population template. The registration method was a diffeomorphic technique from the Advanced Normalization Tools (Ants) package [10]. All registered images were assessed visually for obvious unrealistic deformation or clipping.

2.5. Group difference and statistical normalisation

The images for the two groups of subjects (WS and BS) were averaged independently after the previous step, to yield the average WS and BS templates. Fig 2 illustrates the results, including the difference between the two (as a difference map), revealing some areas with higher and lower intensity between the two groups. However, since the X-ray intensity is relative, it is challenging to draw any conclusion. We thus proceeded to normalise and quantify statistically such differences.

We randomly created 2 groups of images with exactly half WS and half BS images, and computed the difference. Under the null hypothesis, that there is no intensity difference between WS and BS, this should result in a flat image. To account for expected variation, we repeated this 100

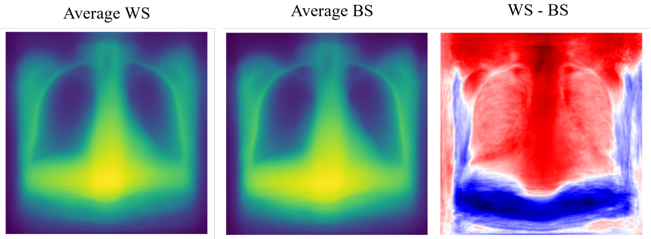

Fig. 2. Average of the WS scans (left), BS scans (centre), and the difference between the two (right).

times and computed for each pixel the mean I_μ and standard deviation I_σ of the 100 difference maps. Finally, the difference map between WS and BS $D_{W/S}$ was normalised: $\tilde{D}_{W/S} = (D_{W/S} - I_\mu)/I_\sigma$. As a result, each pixel can now be interpreted as a z-score against the null hypothesis. The results are shown in Fig 3, revealing significant differences between WS and BS (more than 5 standard deviations) around the sternum and the shoulders.

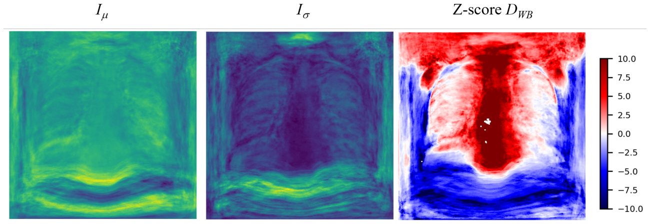

Fig. 3. Mean (left), standard deviation (centre) of difference maps of 100 random groupings with 50% WS and 50% BS. The difference map between WS and BS normalised as a z-score is shown on the right.

2.6. Population atlas of saliency maps

A second approach using saliency maps could then be used, leveraging the same population template technique as described above.

A model with a ResNet34 backbone was trained to classify self-reported race, yielding an area under the curve of 0.97 and 0.97 for BS and WS, respectively. More details about the data, similar model, and the results about race classification can be found in Gichoya et al. [2].

Saliency maps were computed using the 4 methods described in Section 2.2. The non-rigid deformation described in Section 2.4 resulted in a diffeomorphic deformation field that could be inverted to deform the saliency map of each subject, originally computed in the native image space, to the template space. Saliency maps for each group could then be averaged to yield group specific saliency (Fig 4).

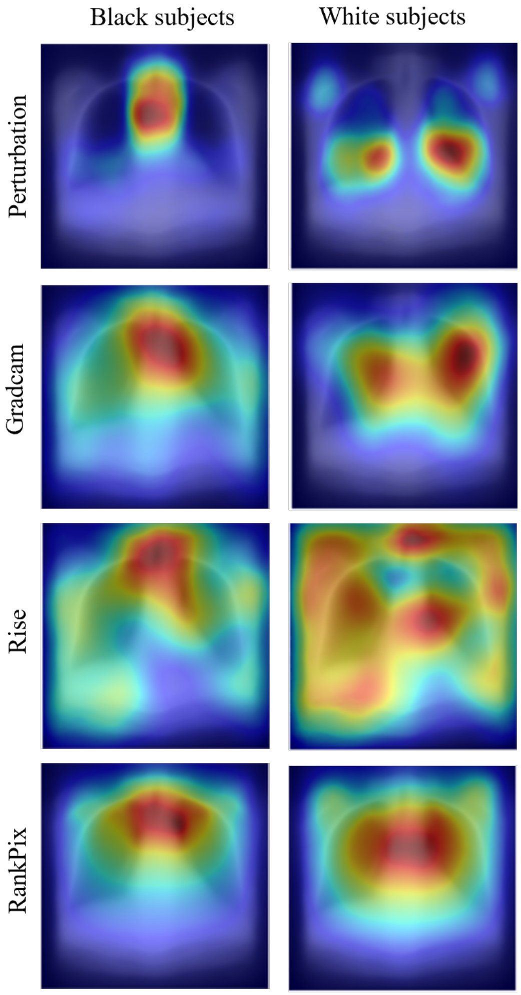

Fig. 4. Average of saliency maps for black subjects (left) and white subjects (right) for the 4 saliency methods.

3. RESULTS AND DISCUSSION

We have described two approaches to investigate where racial information is located within a chest X-ray. The first method showed the calibrated difference between the 2 group templates, whereas the second method averages saliency maps per group. The results are consistent but also present some differences. Indeed, both approaches suggest that the X-ray intensity of the sternum and clavicle areas are different (more dense for white subjects).

Saliency maps are prevalent techniques to explain a deep learning model output. Various methods exist, and we tested 4 of the more promising ones. All aim to "explain" the model output. It is challenging to compare different saliency maps as no ground truth is available, and all optimise the map differently. We first qualitatively compared saliency methods by excluding any X-ray that showed devices and wires from the dataset during training. During testing on the X-ray that had been removed, we expected the saliency map to highlight the abnormal objects. They all did, but with various usefulness. Extremal Perturbation tended to provide a circular area and was challenged when multiple objects were present (Fig 1, middle column). The RISE method resulted in diffuse and mostly uninformative maps. Gradcam and Rankpix provided the most informative results, with Rankpix providing sharper and more defined areas.

We also described a method comparing atlases between the two groups. Averaging X-rays of each group and looking at their difference ought to show where the intensity is different. This kind of approach is routinely used for brain analysis, such as voxel brain morphometry (VBM) [11]. However, in VBM, the average of segmentations is analyzed (instead of the intensity), which requires compensating for the anatomical deformation introduced by the registration. For chest X-rays, the X-ray absorption (x-ray intensity) was investigated instead of the anatomical differences, which presented too much dissimilitude to be meaningful in our experiments. Further analysis of the anatomical variation between groups using careful non-rigid registration will be the subject of future work.

Comparing group intensity averages could be achieved in a statistically meaningful way by bootstrapping the null hypothesis to compute the z-score for each pixel. We did not correct for multiple comparisons because the results were very large homogeneous areas (as opposed to noisy or patchy output).

The results using two different approaches are consistent and point to the sternum and clavicle as candidate sources of information. The actual physiological, anatomical, or confounding reasons to account for this difference remains to be elucidated, but our work present an important first step towards this goal.

4. ETHICS AND ACKNOWLEDGEMENT

This research study was conducted retrospectively using human subject data made available in open access (see section2.1). Further approval was secured through CSIRO ethical board. This work was funded in part through an Australian Department of Industry, Energy and Resources CRC-P project between CSIRO, Maxwell Plus and I-Med Radiology Network.

5. REFERENCES

[1] Jarrel C. Y. Seah, Cyril H. M. Tang, Quinlan D. Buchlak, Xavier G. Holt, Jeffrey B. Wardman, Anuar Aimoldin, Nazanin Esmaili, Hassan Ahmad, Hung Pham, John F. Lambert, Ben Hachey, Stephen J. F. Hogg, Benjamin P. Johnston, Christine Bennett, Luke Oakden-Rayner, Peter Brotchie, and Catherine M. Jones, "Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study," *The Lancet Digital Health*, vol. 3, no. 8, pp. e496–e506, Aug. 2021, Publisher: Elsevier.

[2] Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al., "Ai recognition of patient race in medical imaging: a modelling study," *The Lancet Digital Health*, 2022.

[3] RSNA Pneumonia Detection, "Kaggle," https://www.kaggle.com/c/rsna- pneumonia-detection-challenge, 2018, Accessed: August 2019.

[4] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[5] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, vol. 6, no. 1, pp. 1–8, 2019.

[6] Ruth Fong, Mandela Patrick, and Andrea Vedaldi, "Understanding deep networks via extremal perturbations and smooth masks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2950–2958.

[7] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[8] Vitali Petsiuk, Abir Das, and Kate Saenko, "Rise: Randomized input sampling for explanation of black-box models," *arXiv preprint arXiv:1806.07421*, 2018.

[9] Salamata Konate, Léo Lebrat, Rodrigo Santa Cruz, Clinton Fookes, Andrew Bradley, and Olivier Salvado, "Bias identification with rankpix saliency," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] Brian B Avants, Nick Tustison, Gang Song, et al., "Advanced normalization tools (ants)," *Insight j*, vol. 2, no. 365, pp. 1–35, 2009.

[11] J. Ashburner and K.J. Friston, "Voxel based morphometry," in *Encyclopedia of Neuroscience*, Larry R. Squire, Ed., pp. 471–477. Academic Press, Oxford, 2009.