On the Impact of Gender Bias in Medical Imaging Classifiers for Computer-aided Diagnosis

Anonymous Author(s)
Affiliation
Address
email

1 Research Problem

Artificial intelligence (AI) influences almost every aspect of our daily life. In particular, the rise of AI in healthcare during the last few years is changing the way medical doctors diagnose, especially when dealing with medical images. AI systems can not only augment the information provided by such images with useful annotations [1], but they are also taking autonomous decisions by performing computer assisted diagnosis (CAD) [2, 3]. In this context, care must be taken in the risk associated with error and misbehaviour of algorithms, especially since their decision comprises a delicate domain such as health care.

Although the interest in performing fair and unbiased evaluations of AI medical systems exists since the 80's [4], the ethical and moral aspects of AI have gained relevance in the last few years, showing that human bias, such as gender and race bias may be inherited by AI systems in multiple contexts [5, 6]. In the last years, the research community of gendered innovations [7] has worked to create awareness and integrate sex and gender analysis into all phases of basic and applied research. However, to date, there is no study reflecting such analysis in the context of medical imaging and computer assisted diagnosis.

In this work, we perform the first large-scale study that quantifies the influence of gender imbalance in medical imaging datasets used to train AI-based CAD systems. We employ a classification model based on deep neural networks, which achieves state-of-the-art results when diagnosing 14 common thoracic diseases using X-ray images [8]. We analyze the performance over male and female patients when the model is trained with different gender unbalance ratios, providing empirical evidence about the bias acquired by such systems.

2 Experiments

We use a frontal view chest X-ray image dataset, that contains over 100.000 images with 14 different common pathologies [8]. The dataset has 63340 images of male patients and 48780 female patients. All images were automatically labeled using natural language processing tools to extract such information from radiology reports (see [9] for a detailed method description).

Given a frontal X-ray image, we trained CAD models to predict the presence or absence of the 14 thoracic diseases considering male-only and female-only training datasets. We then evaluated both classifiers in male and female patients separately, and reported their performance using the well-known area under the receiving operating curve (AUC) [10]. Since we are focusing on the effect of gender imbalance, we guaranteed by construction that male and female folds include the same number of pathological cases per class to avoid other sources of bias.

A Densely Connected Convolutional Neural Network (DenseNet) [11] architecture with 14 outputs representing the probability of each disease was used for classification. We adopted a Keras implementation of the DenseNet-121 (publicly available at: *https://github.com/brucechou1983/CheXNet-Keras*) which has shown to achieve state-of-the-art results in X-ray image classification [8]. The network has 121 convolutional layers and a final fully connected layer producing a 14-dimensional output, after which we apply an element-wise sigmoid non-linearity. A model pre-trained on ImageNet [12] was used to initialize the network. We trained it end-to-end using Adam optimizer with standard parameters ($\beta1$=0.9 and $\beta2$=0.999), a batch size of 32 and an initial learning rate of 0.001 that was decayed by a factor of 10 each time the validation loss plateaus after an epoch.

For every experiment, 20 models were trained using random sampling and the area under the curve (AUC) was used to measure classification performance. For each random subset, male and female patients were evaluated separately, considering for each gender 20% test, 70% training and 10% validation, ensuring that images from one patient were not overlapping in different splits. In all cases, we tested each model with a female test set and male test set independently.

3 Results and discussion

The experimental results shown in Figure 1 show a consistent decrease in performance when using male patients for training and female for testing, and viceversa. In particular, we found that 16 out of 28 cases (14 diseases per gender) present a significant decrease in performance, according to a Mann–Whitney U test, considering a p-value < 0.01.

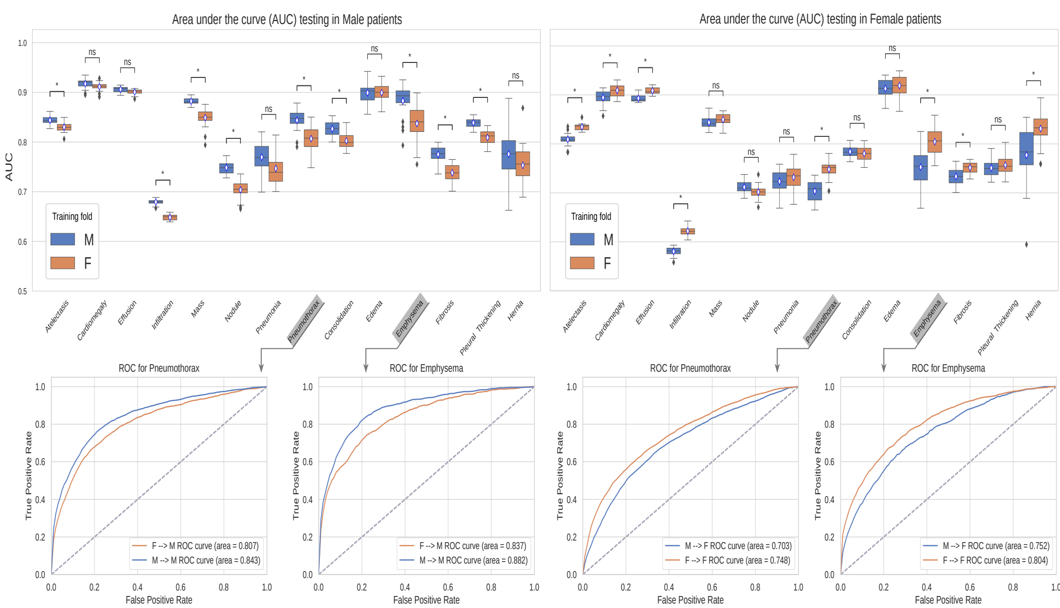

Figure 1: Top: AUC for each disease corresponding to a CNN trained only with male patients (blue) and trained only with female patients (orange). Bottom: ROC curve from specific diseases (p-value < 0.01 according to a Mann–Whitney U test).

4 Conclusions

We have shown how algorithms can present gender bias if they are trained in unbalanced data sets, producing serious misbehaviour for the under-represented class. This raises the alarm for national agencies in charge of regulating and approving CAD systems, which should include explicit gender balance and diversity recommendations. We also establish a new open problem for the academic medical image computing community which needs to be addressed by novel algorithms endowed with robustness to gender imbalance.

In the future we plan to develop algorithmic solutions to tackle gender bias problems in cases where it is difficult to obtain balanced datasets with annotations. We plan to employ adversarial domain adaptation techniques [13] to train classification models which are invariant to patient gender.

References

[1] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[2] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[3] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342, 2018.

[4] B Chandrasekaran. On evaluating artificial intelligence systems for medical diagnosis. *AI magazine*, 4(2):34–34, 1983.

[5] James Zou and Londa Schiebinger. Ai can be sexist and racist—it's time to make it fair, 2018.

[6] Matthew Hutson et al. Even artificial intelligence can acquire biases against race and gender. *Science Magazine*, 10, 2017.

[7] Londa Schiebinger and Martina Schraudner. Interdisciplinary approaches to achieving gendered innovations in science, medicine, and engineering1. *Interdisciplinary Science Reviews*, 36(2):154–167, 2011.

[8] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[9] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[10] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.