Medical Xpress

Study reveals why AI models that analyze medical images can be biased

June 28 2024, by Anne Trafton

Credit: Unsplash/CC0 Public Domain

Artificial intelligence models often play a role in medical diagnoses, especially when it comes to analyzing images such as X-rays. However, studies have found that these models don't always perform well across all demographic groups, usually faring worse in women and people of color.

These models have also been shown to develop some surprising abilities. In 2022, MIT researchers reported that AI models can make accurate predictions about a patient's race from their chest X-rays—something that the most skilled radiologists can't do.

That research team has now found that the models that are most accurate at making demographic predictions also show the biggest "fairness gaps"—that is, discrepancies in their ability to accurately diagnose images of people of different races or genders. The findings suggest that these models may be using "demographic shortcuts" when making their diagnostic evaluations, which lead to incorrect results for women, Black people, and other groups, the researchers say.

"It's well-established that high-capacity machine-learning models are good predictors of human demographics such as self-reported race or sex or age. This paper re-demonstrates that capacity, and then links that capacity to the lack of performance across different groups, which has never been done," says Marzyeh Ghassemi, an MIT associate professor of electrical engineering and computer science, a member of MIT's Institute for Medical Engineering and Science, and the senior author of the study.

The researchers also found that they could retrain the models in a way that improves their fairness. However, their approach to "debiasing" worked best when the models were tested on the same types of patients on whom they were trained, such as patients from the same hospital. When these models were applied to patients from different hospitals, the fairness gaps reappeared.

"I think the main takeaways are, first, you should thoroughly evaluate any external models on your own data because any fairness guarantees that model developers provide on their training data may not transfer to your population. Second, whenever sufficient data is available, you should train models on your own data," says Haoran Zhang, an MIT graduate student and one of the lead authors of the new paper.

MIT graduate student Yuzhe Yang is also a lead author of the paper, which will appear in *Nature Medicine*. Judy Gichoya, an associate professor of radiology and imaging sciences at Emory University School of Medicine, and Dina Katabi, the Thuan and Nicole Pham Professor of Electrical Engineering and Computer Science at MIT, are also authors of the paper.

Removing bias

As of May 2024, the FDA has approved 882 AI-enabled medical devices, with 671 of them designed to be used in radiology. Since 2022, when Ghassemi and her colleagues showed that these diagnostic models can accurately predict race, they and other researchers have shown that such models are also very good at predicting gender and age, even though the models are not trained on those tasks.

"Many popular machine learning models have superhuman demographic prediction capacity—radiologists cannot detect self-reported race from a chest X-ray," Ghassemi says. "These are models that are good at predicting disease, but during training are learning to predict other things that may not be desirable."

In this study, the researchers set out to explore why these models don't work as well for certain groups. In particular, they wanted to see if the models were using demographic shortcuts to make predictions that ended up being less accurate for some groups. These shortcuts can arise in AI models when they use demographic attributes to determine whether a medical condition is present, instead of relying on other features of the images.

Using publicly available chest X-ray datasets from Beth Israel Deaconess Medical Center in Boston, the researchers trained models to predict whether patients had one of three different medical conditions: fluid buildup in the lungs, collapsed lung, or enlargement of the heart. Then, they tested the models on X-rays that were held out from the training data.

Overall, the models performed well, but most of them displayed "fairness gaps"—that is, discrepancies between accuracy rates for men and women, and for white and Black patients.

The models were also able to predict the gender, race, and age of the X-ray subjects. Additionally, there was a significant correlation between each model's accuracy in making demographic predictions and the size of its fairness gap. This suggests that the models may be using demographic categorizations as a shortcut to make their disease predictions.

The researchers then tried to reduce the fairness gaps using two types of strategies. For one set of models, they trained them to optimize "subgroup robustness," meaning that the models are rewarded for having better performance on the subgroup for which they have the worst performance, and penalized if their error rate for one group is higher than the others.

In another set of models, the researchers forced them to remove any demographic information from the images, using "group adversarial" approaches. Both of these strategies worked fairly well, the researchers found.

"For in-distribution data, you can use existing state-of-the-art methods to reduce fairness gaps without making significant trade-offs in overall performance," Ghassemi says. "Subgroup robustness methods force models to be sensitive to mispredicting a specific group, and group adversarial methods try to remove group information completely."

Not always fairer

However, those approaches only worked when the models were tested on data from the same types of patients that they were trained on—for example, only patients from the Beth Israel Deaconess Medical Center dataset.

When the researchers tested the models that had been "debiased" using the BIDMC data to analyze patients from five other hospital datasets, they found that the models' overall accuracy remained high, but some of them exhibited large fairness gaps.

"If you debias the model in one set of patients, that fairness does not necessarily hold as you move to a new set of patients from a different hospital in a different location," Zhang says.

This is worrisome because in many cases, hospitals use models that have been developed on data from other hospitals, especially in cases where an off-the-shelf model is purchased, the researchers say.

"We found that even state-of-the-art models which are optimally performant in data similar to their training sets are not optimal—that is, they do not make the best trade-off between overall and subgroup performance—in novel settings," Ghassemi says. "Unfortunately, this is actually how a model is likely to be deployed. Most models are trained and validated with data from one hospital, or one source, and then deployed widely."

The researchers found that the models that were debiased using group adversarial approaches showed slightly more fairness when tested on new patient groups than those debiased with subgroup robustness methods. They now plan to try to develop and test additional methods to see if they can create models that do a better job of making fair predictions on new datasets.

The findings suggest that hospitals that use these types of AI models should evaluate them on their own patient population before beginning to use them, to make sure they aren't giving inaccurate results for certain groups.

More information: The limits of fair medical imaging AI in real-world generalization, *Nature Medicine* (2024). DOI: 10.1038/s41591-024-03113-4

Provided by Massachusetts Institute of Technology

Citation: Study reveals why AI models that analyze medical images can be biased (2024, June 28) retrieved 30 June 2024 from https://medicalxpress.com/news/2024-06-reveals-ai-medical-images-biased.html