TechXplore

Medical AI models rely on 'shortcuts' that could lead to misdiagnosis of COVID-19

May 31 2021

Credit: Unsplash/CC0 Public Domain

Artificial intelligence promises to be a powerful tool for improving the speed and accuracy of medical decision-making to improve patient outcomes. From diagnosing disease, to personalizing treatment, to predicting complications from surgery, AI could become as integral to patient care in the future as imaging and laboratory tests are today.

But as University of Washington researchers discovered, AI models—like humans—have a tendency to look for shortcuts. In the case of AI-assisted disease detection, these shortcuts could lead to diagnostic errors if deployed in clinical settings.

In a new paper published May 31 in *Nature Machine Intelligence*, UW researchers examined multiple models recently put forward as potential tools for accurately detecting COVID-19 from chest radiography, otherwise known as chest X-rays. The team found that, rather than learning genuine medical pathology, these models rely instead on shortcut learning to draw spurious associations between medically irrelevant factors and disease status. Here, the models ignored clinically significant indicators and relied instead on characteristics such as text markers or patient positioning that were specific to each dataset to predict whether someone had COVID-19.

"A physician would generally expect a finding of COVID-19 from an X-ray to be based on specific patterns in the image that reflect disease processes," said co-lead author Alex DeGrave, who is pursuing his doctorate in the Paul G. Allen School of Computer Science & Engineering and a medical degree as part of the UW's Medical Scientist Training Program. "But rather than relying on those patterns, a system using shortcut learning might, for example, judge that someone is elderly and thus infer that they are more likely to have the disease because it is more common in older patients. The shortcut is not wrong per se, but the association is unexpected and not transparent. And that could lead to an inappropriate diagnosis."

Shortcut learning is less robust than genuine medical pathology and usually means the model will not generalize well outside of the original setting, the team said.

"A model that relies on shortcuts will often only work in the hospital in which it was developed, so when you take the system to a new hospital, it fails—and that failure can point doctors toward the wrong diagnosis and improper treatment," DeGrave said.

Combine that lack of robustness with the typical opacity of AI decision-making, and such a tool could go from a potential life-saver to a liability.

The lack of transparency is one of the factors that led the team to focus on explainable AI techniques for medicine and science. Most AI is regarded as a "black box"—the model is trained on massive datasets and it spits out predictions without anyone knowing precisely how the model came up with a given result. With explainable AI, researchers and practitioners are able to understand, in detail, how various inputs and their weights contributed to a model's output.

The team used these same techniques to evaluate the trustworthiness of models recently touted for appearing to accurately identify cases of COVID-19 from chest X-rays. Despite a number of published papers heralding the results, the researchers suspected that something else may have been happening inside the black box that led to the models' predictions.

Specifically, the team reasoned that these models would be prone to a condition known as "worst-case confounding," owing to the lack of training data available for such a new disease. This scenario increased the likelihood that the models would rely on shortcuts rather than learning the underlying pathology of the disease from the training data.

"Worst-case confounding is what allows an AI system to just learn to recognize datasets instead of learning any true disease pathology," said co-lead author Joseph Janizek, who is also a doctoral student in the Allen School and earning a medical degree at the UW. "It's what happens when all of the COVID-19 positive cases come from a single dataset while all of the negative cases are in another. And while researchers have come up with techniques to mitigate associations like this in cases where those associations are less severe, these techniques don't work in situations where you have a perfect association between an outcome such as COVID-19 status and a factor like the data source."

The team trained multiple deep convolutional neural networks on X-ray images from a dataset that replicated the approach used in the published papers. First they tested each model's performance on an internal set of images from that initial dataset that had been withheld from the training data. Then the researchers tested how well the models performed on a second, external dataset meant to represent new hospital systems.

While the models maintained their high performance when tested on images from the internal dataset, their accuracy was reduced by half on the second set. The researchers referred to this as a "generalization gap" and cited it as strong evidence that confounding factors were responsible for the models' predictive success on the initial dataset.

The team then applied explainable AI techniques, including generative adversarial networks and saliency maps, to identify which image features were most important in determining the models' predictions.

The researchers trained the models on a second dataset, which contained positive and negative COVID-19 cases drawn from similar sources, and was therefore presumed to be less prone to confounding. But even those models exhibited a corresponding drop in performance when tested on external data.

These results upend the conventional wisdom that confounding poses less of an issue when datasets are derived from similar sources. They also reveal the extent to which high-performance medical AI systems could exploit undesirable shortcuts rather than the desired signals.

"My team and I are still optimistic about the clinical viability of AI for medical imaging. I believe we will eventually have reliable ways to prevent AI from learning shortcuts, but it's going to take some more work to get there," said senior author Su-In Lee, a professor in the Allen School. "Going forward, explainable AI is going to be an essential tool for ensuring these models can be used safely and effectively to augment medical decision-making and achieve better outcomes for patients."

Despite the concerns raised by the team's findings, it is unlikely that the models the team studied have been deployed widely in the clinical setting, DeGrave said. While there is evidence that at least one of the faulty models—COVID-Net—was deployed in multiple hospitals, it is unclear whether it was used for clinical purposes or solely for research.

"Complete information about where and how these models have been deployed is unavailable, but it's safe to assume that clinical use of these models is rare or nonexistent," DeGrave said. "Most of the time, healthcare providers diagnose COVID-19 using a laboratory test, PCR, rather than relying on chest radiographs. And hospitals are averse to liability, making it even less likely that they would rely on a relatively untested AI system."

Researchers looking to apply AI to disease detection will need to revamp their approach before such models can be used to make actual treatment decisions for patients, Janizek said.

"Our findings point to the importance of applying explainable AI techniques to rigorously audit medical AI systems," Janizek said. "If you look at a handful of X-rays, the AI system might appear to behave well. Problems only become clear once you look at many images. Until we have methods to more efficiently audit these systems using a greater sample size, a more systematic application of explainable AI could help researchers avoid some of the pitfalls we identified with the COVID-19 models."

This group has already demonstrated the value of explainable AI for a range of medical applications beyond imaging. These include tools for assessing patient risk factors for complications during surgery and targeting cancer therapies based on an individual's molecular profile.

More information: AI for radiographic COVID-19 detection selects shortcuts over signal, *Nature Machine Intelligence* (2021). DOI: 10.1038/s42256-021-00338-7 , www.nature.com/articles/s42256-021-00338-7

Provided by University of Washington

Citation: Medical AI models rely on 'shortcuts' that could lead to misdiagnosis of COVID-19 (2021, May 31) retrieved 27 March 2023 from https://techxplore.com/news/2021-05-medical-ai-shortcuts-misdiagnosis-covid-.html