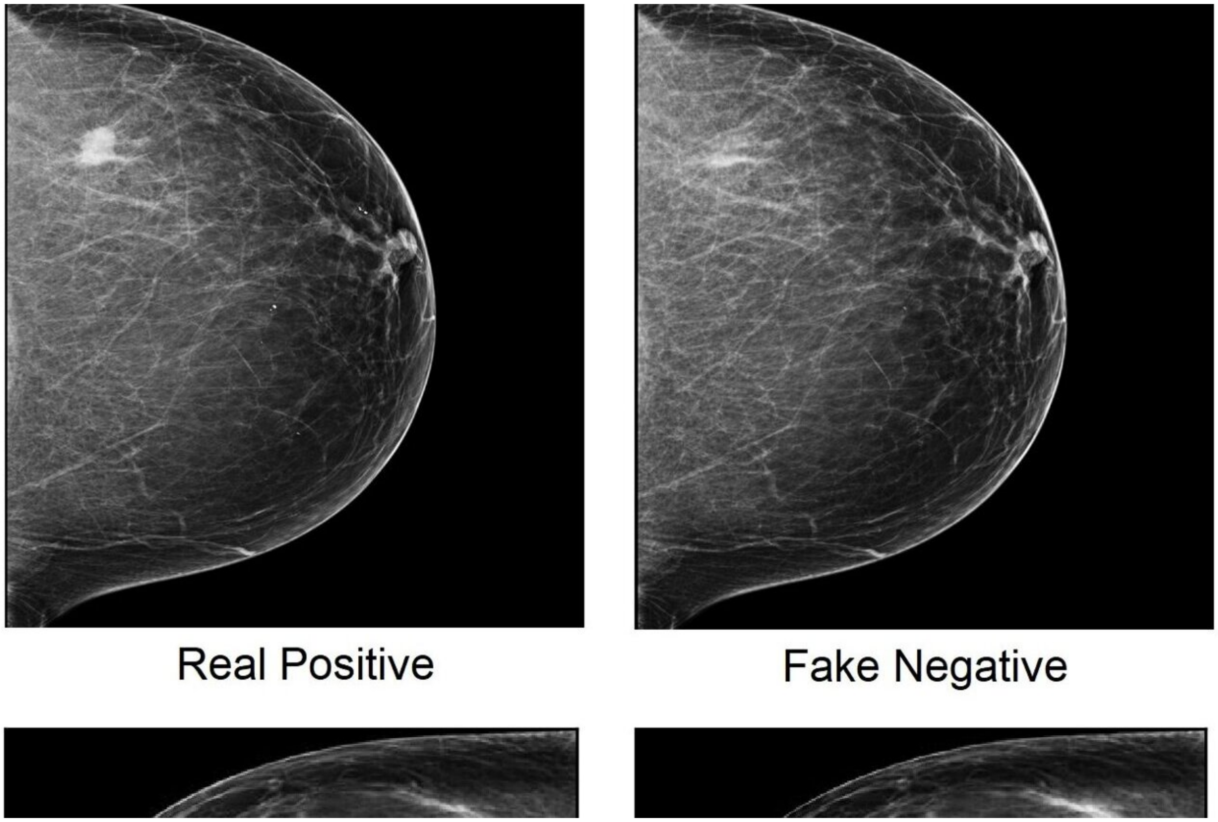
MedicalXpress

Cancer-spotting AI and human experts can be fooled by image-tampering attacks

December 14 2021



Mammogram images showing real cancer-positive (top left) and cancer-negative (bottom left) cases, with cancerous tissue indicated by white spot. A 'generative adversarial network' program removed cancerous regions from the cancer-positive image, creating a fake negative image (top right) and inserted cancerous regions to the cancer-negative image, creating a fake positive (bottom right). Credit: Q. Zhou et al., Nat. Commun. 2021

Artificial intelligence (AI) models that evaluate medical images have potential to speed up and improve accuracy of cancer diagnoses, but they may also be vulnerable to cyberattacks. In a new study, University of Pittsburgh researchers simulated an attack that falsified mammogram images, fooling both an AI breast cancer diagnosis model and human breast imaging radiologist experts.

The study, published today in *Nature Communications*, brings attention to a potential safety issue for medical AI known as "adversarial attacks," which seek to alter images or other inputs to make models arrive at incorrect conclusions.

"What we want to show with this study is that this type of attack is possible, and it could lead AI models to make the wrong diagnosis—which is a big patient safety issue," said senior author Shandong Wu, Ph.D., associate professor of radiology, biomedical informatics and bioengineering at Pitt. "By understanding how AI models behave under adversarial attacks in medical contexts, we can start thinking about ways to make these models safer and more robust."

AI-based image recognition technology for cancer detection has advanced rapidly in recent years, and several breast cancer models have U.S. Food and Drug Administration (FDA) approval. According to Wu, these tools can rapidly screen mammogram images and identify those most likely to be cancerous, helping radiologists be more efficient and accurate.

But such technologies are also at risk from cyberthreats, such as adversarial attacks. Potential motivations for such attacks include insurance fraud from health care providers looking to boost revenue or companies trying to adjust clinical trial outcomes in their favor. Adversarial attacks on medical images range from tiny manipulations that change the AI's decision, but are imperceptible to the human eye, to more sophisticated versions that target sensitive contents of the image, such as cancerous regions —making them more likely to fool a human.

To understand how AI would behave under this more complex type of adversarial attack, Wu and his team used mammogram images to develop a model for detecting breast cancer. First, the researchers trained a deep learning algorithm to distinguish cancerous and benign cases with more than 80% accuracy. Next, they developed a so-called "generative adversarial network" (GAN)—a computer program that generates false images by inserting or removing cancerous regions from negative or positive images, respectively, and then they tested how the model classified these adversarial images.

Of 44 positive images made to look negative by the GAN, 42 were classified as negative by the model, and of 319 negative images made to look positive, 209 were classified as positive. In all, the model was fooled by 69.1% of the fake images.

In the second part of the experiment, the researchers asked five human radiologists to distinguish whether mammogram images were real or fake. The experts accurately identified the images' authenticity with accuracy of between 29% and 71%, depending on the individual.

"Certain fake images that fool AI may be easily spotted by radiologists. However, many of the adversarial images in this study not only fooled the model, but they also fooled experienced human readers," said Wu, who is also the director of the Intelligent Computing for Clinical Imaging Lab and the Pittsburgh Center for AI Innovation in Medical Imaging. "Such attacks could potentially be very harmful to patients if they lead to an incorrect cancer diagnosis."

According to Wu, the next step is developing ways to make AI models more robust to adversarial attacks.

"One direction that we are exploring is 'adversarial training' for the AI model," he explained. "This involves pre-generating adversarial images and teaching the model that these images are manipulated."

With the prospect of AI being introduced to medical infrastructure, Wu said that cybersecurity education is also important to ensure that hospital technology systems and personnel are aware of potential threats and have technical solutions to protect patient data and block malware.

"We hope that this research gets people thinking about medical AI model safety and what we can do to defend against potential attacks, ensuring AI systems function safely to improve patient care," he added.

Other authors who contributed to the study were Qianwei Zhou, Ph.D., of Pitt and Zhejiang University of Technology in China; Margarita Zuley, M.D., Bronwyn Nair, M.D., Adrienne Vargo, M.D., Suzanne Ghannam, M.D., and Dooman Arefan, Ph.D., all of Pitt and UPMC; Yuan Guo, M.D., of Pitt and Guangzhou First People's Hospital in China; Lu Yang, M.D., of Pitt and Chongqing University Cancer Hospital in China.

More information: A machine and human reader study on AI diagnosis model safety under attacks of adversarial images, *Nature Communications* (2021). DOI: 10.1038/s41467-021-27577-x

Provided by University of Pittsburgh

Citation: Cancer-spotting AI and human experts can be fooled by image-tampering attacks (2021, December 14) retrieved 9 April 2024 from https://medicalxpress.com/news/2021-12-cancer-spotting-ai-human-experts-image-tampering.html