time.com

AI Can Now Make Medical Predictions from Raw Data Through 'Deep Learning.' But Can it Be Trusted?

By Eric Bender / Undark

14–18 minutes

—In clinics around the world, a type of artificial intelligence called deep learning is starting to supplement or replace humans in common tasks such as analyzing medical images. Already, at Massachusetts General Hospital in Boston, “every one of the 50,000 screening mammograms we do every year is processed through our deep learning model, and that information is provided to the radiologist,” says Constance Lehman, chief of the hospital’s breast imaging division.

In deep learning, a subset of a type of artificial intelligence called machine learning, computer models essentially teach themselves to make predictions from large sets of data. The raw power of the technology has improved dramatically in recent years, and it’s now used in everything from medical diagnostics to online shopping to autonomous vehicles.

But deep learning tools also raise worrying questions because they solve problems in ways that humans can’t always follow. If the connection between the data you feed into the model and the

output it delivers is inscrutable — hidden inside a so-called black box — how can it be trusted? Among researchers, there's a growing call to clarify how deep learning tools make decisions — and a debate over what such interpretability might demand and when it's truly needed. The stakes are particularly high in medicine, where lives will be on the line.

Still, the potential benefits are clear. In Mass General's mammography program, for instance, the current deep learning model helps detect dense breast tissue, a risk factor for cancer. And Lehman and Regina Barzilay, a computer scientist at the Massachusetts Institute of Technology, have created another deep learning model to predict a woman's risk of developing breast cancer over five years — a crucial component of planning her care. In a 2019 retrospective study of mammograms from about 40,000 women, the researchers found the deep learning system substantially outperformed the current gold-standard approach on a test set of about 4,000 of these women. Now undergoing further testing, the new model may enter routine clinical practice at the hospital.

As for the debate about whether humans can really understand deep learning systems, Barzilay sits firmly in the camp that it's possible. She calls the black box problem "a myth."

One part of the myth, she says, is that deep learning systems can't explain their results. But "there are lots of methods in machine language that allow you to interpret the results," she says. Another part of the myth, in her opinion, is that doctors have to understand how the system makes its decision in order to use it. But medicine is crammed with advanced technologies that work in ways that

clinicians really don't understand — for instance, the magnetic resonance imaging (MRI) that gathers the mammography data to begin with.

That doesn't answer the concerns of all physicians. Many machine learning tools are still black boxes "that render verdicts without any accompanying justification," notes a group of physicians and researchers in a recent paper in BMJ Clinical Research. "Many think that, as a new technology, the burden of proof is on machine learning to account for its predictions," the paper's authors continue. "If doctors do not understand why the algorithm made a diagnosis, then why should patients trust the recommended course of treatment?"

And among computer scientists who study machine learning, "this discussion of interpretability has gone completely off the rails," says Zachary Lipton, a computer scientist at Carnegie Mellon University. Often, models offered for interpretability simply don't work well, he says, and there's confusion about what the systems actually deliver.

"We have people in the field who are able to turn the crank but don't actually know what they're doing," he adds, "and don't actually understand the foundational underpinnings of what they're doing."

Deep learning tools build on the concept of neural networks, originally inspired by the human brain and composed of nodes that act somewhat like brain cells. Deep learning models assemble multiple layers of these artificial neurons into a vast web of evolving connections. And the models juggle data on levels far

beyond what the human mind can follow.

Understanding how the models work matters in some applications more than others. Worries about whether Amazon is offering perfect suggestions for your aunt's birthday gift aren't the same, for example, as worries about the trustworthiness of the tools your doctor is using to detect tumors or oncoming heart attacks.

Computer scientists are trying many approaches to make deep learning less opaque, at least to their peers. A model of breast cancer risk, for example, can use a heat map approach, letting radiologists zoom into areas of the mammography image that the model pays attention to when it makes a prediction. The model can then extract and highlight snippets of text that describe what it sees.

Deep learning models can also present images of other regions that are similar to these targeted areas, and human experts can then assess the machine's choices. Another popular technique applies math that is more immediately understandable to subsets of the data to approximate how the deep learning model is handling the full dataset.

"We will learn more about what explanations are convincing to humans when these models are integrated into care, and we can see how the human mind can help to control and validate their predictions," Barzilay says.

In London, a team from Moorfields Eye Hospital and DeepMind, a subsidiary of Google parent company Alphabet, also seeks to deliver explanations in depth. They have used deep learning to triage scans of patient eyes. The system takes in three-

dimensional eye scans, analyzes them, and picks cases that need urgent referral — and it works as well as or better than human experts. The model gives and rates several possible explanations for each diagnosis and shows how it has labeled the parts of the patient's eye.

As a general strategy in bringing deep learning to the clinic, "the key is to build the best system but then analyze its behavior," says Anna Goldenberg, a senior scientist in genetics and genome biology at SickKids Research Institute in Toronto, who is partnering with clinicians to build a model that can predict cardiac arrests. "I think we want both. I think it's achievable."

Models like Mass General's and Moorfields' are well-designed, with doctor input and clinical results in peer-reviewed scientific publications, and they rest on solid technical foundations. But few attempts at interpretability will make it this far, Lipton says.

More often, such interpretations don't show a real connection between the data that go in and what comes out. "Basically people have been looking at the pretty pictures and picking the one that looks like what they wanted to see in the first place," Lipton adds. "Increasingly, you wind up with people just throwing spaghetti at the wall and calling it explanations."

Even if computer scientists find a way to show how a deep learning tool works, doctors will have the final say on whether the explanations are sufficient. Doctors aren't just interested in theoretical accuracy — they need to know the system works in the real world.

For example, when doctors are trying to spot a small tumor or early

signs of an upcoming cardiac arrest, “false positives are not so problematic, because clinicians try to avoid detecting things late,” says Goldenberg. “But false negatives are a really big problem.” If the rate of false positives is too high, however, then doctors may not pay attention to the system at all.

When physicians see the clinical factors that are considered in the deep learning system, it’s easier for them to interpret the results. “Without understanding that, they’re suspicious,” Goldenberg says. “They don’t need to understand exactly how the system works or how deep learning works. They need to understand how the system would make a decision compared to them. So they will throw some cases against the system and see what it does and then see whether they trust it.”

Deep learning studies should begin by analyzing large numbers of suitable, existing medical records, experts say. In some cases, such as Goldenberg’s cardiac arrest model, she says the next step may be to run a trial where “we can let the system run, getting the real time inputs but not giving any feedback back to the clinician, and seeing the difference between the practice and what our system is predicting.”

“Before we point the finger too much at AI, we should look at all our other practices that are ripe with false positives and false negatives, and all other practices that are black boxes, based on publications that in reality few doctors read in detail,” says Isaac Kohane, a bioinformatician and physician at Harvard Medical School.

Because AI is just coming into practice, it hasn’t seen the same

kind of vetting as some other technologies, Kohane adds. “And because it doesn’t look the same as a blood test or imaging test, the health care system or the regulatory authorities have not yet figured out the right way to make sure that we know it’s acceptably safe, whatever acceptably is.”

Kohane says his biggest concern is that no one really knows how well the new models work. “We should be more worried about what is a false positive rate and what is a false negative rate over time of these programs,” he adds, “so that even if they are used in black box fashion like the rest of medicine, they are sufficiently reliable.”

There is still a lot of work needed to assess the models’ performance. A 2019 study in Lancet Digital Health, which analyzed results from 69 studies on medical deep learning tools, found that the models performed as well as health care professionals. But the paper’s authors also cautioned that few of the studies pitted the models head-to-head against human experts on the same dataset.

Such studies also may miss many subtle but important issues, and they often rest on shaky assumptions, Lipton says. One major problem: A model’s accuracy is only as meaningful as the data on which it is based. “The truth is that you don’t even usually have historical data that is representative of any real-world process,” he says. “You have some kind of weird Frankenstein data that was cobbled together from a bunch of sources.”

Given deep learning’s growing power and the worries it incites, how will medical gatekeepers view the technology?

“The black box story gets a lot of emphasis,” says Eric Topol,

director of the Scripps Research Translational Institute in La Jolla, California. “The question is, will we hold machines hostage to being explainable before they can be used in medicine? I don’t know the answer.”

If prospective trials validate these models, he says, there’s every reason to bring them forward to the clinic with the hope of achieving a happy symbiosis between doctors and machines — even if doctors have no idea why the models work.

“But whenever you lean on machines there’s a pushback, because doctors are used to controlling everything,” Topol says. “When you say: ‘The machine can do this as well as you, or potentially even better,’ there’s a bias against that. Overwhelming proof is one way to deal with that negative bias.”

Descriptions of the issue can carry their own bias. “The term ‘black box’ in itself has a bit of a negative turn,” says Lehman of Mass General. “We don’t know what’s there and we want to fill it with information. But I think the term is unfortunate.” While clinicians may not understand how the tools make every decision, she adds, if they can prove the tools work, “that itself can teach us new explanations. We just have to think differently.”

Today’s medicine abandons vast amounts of data in electronic health records, test tubes, and X-ray machines, Lehman points out. “If we can start to pull out that rich data and be more predictive for who is going to live and who’s going to die, who’s going to need this intervention and who needs that test,” she adds, “we can completely change the entire paradigm of health care.”

The effects will be most profound for people with little or no direct

access to doctors, says Peter Thomas, director of digital innovation at Moorfields Eye Hospital. And Moorfields' eye diagnostic tools, he adds, may help doctors focus on patients who most need their help.

Advocates see adoption coming, and soon. In tasks like analyzing mammograms at Mass General, "deep learning systems are already better than humans," says Barzilay. "Way better than humans."

"Our current models are really complex and they can capture visual patterns which are so subtle that the human eye may not be able to understand them," she adds. "Does it really mean that we shouldn't be benefiting from them because our visual capacity is limited?"

Once deep learning systems clearly demonstrate how well they work, "I don't think that the interpretability will be an issue," Barzilay says. "Many other reasons may prevent adoption, but the black box aspect will not be the main one."

Eric Bender is a science writer based in Boston who primarily covers biomedical research.

This article was originally published on Undark. Read the original article.

- Impeachment Experts Say **Biden Inquiry May Be Weakest in US History**
- **Martin Scorsese Still Has Stories to Tell**
- **Burned Out at Work?** Find Someone to Split Your Job 50-50 With You
- **Jessica Knoll Wants to Correct the Record** on Ted Bundy
- The Most Anticipated Books, Movies, TV, and Music of Fall 2023
- Why It Takes **Forever to Get a Doctor's Appointment**
- Want Weekly Recs on What to Watch, Read, and More? Sign Up for Worth Your Time

  Contact us at letters@time.com.