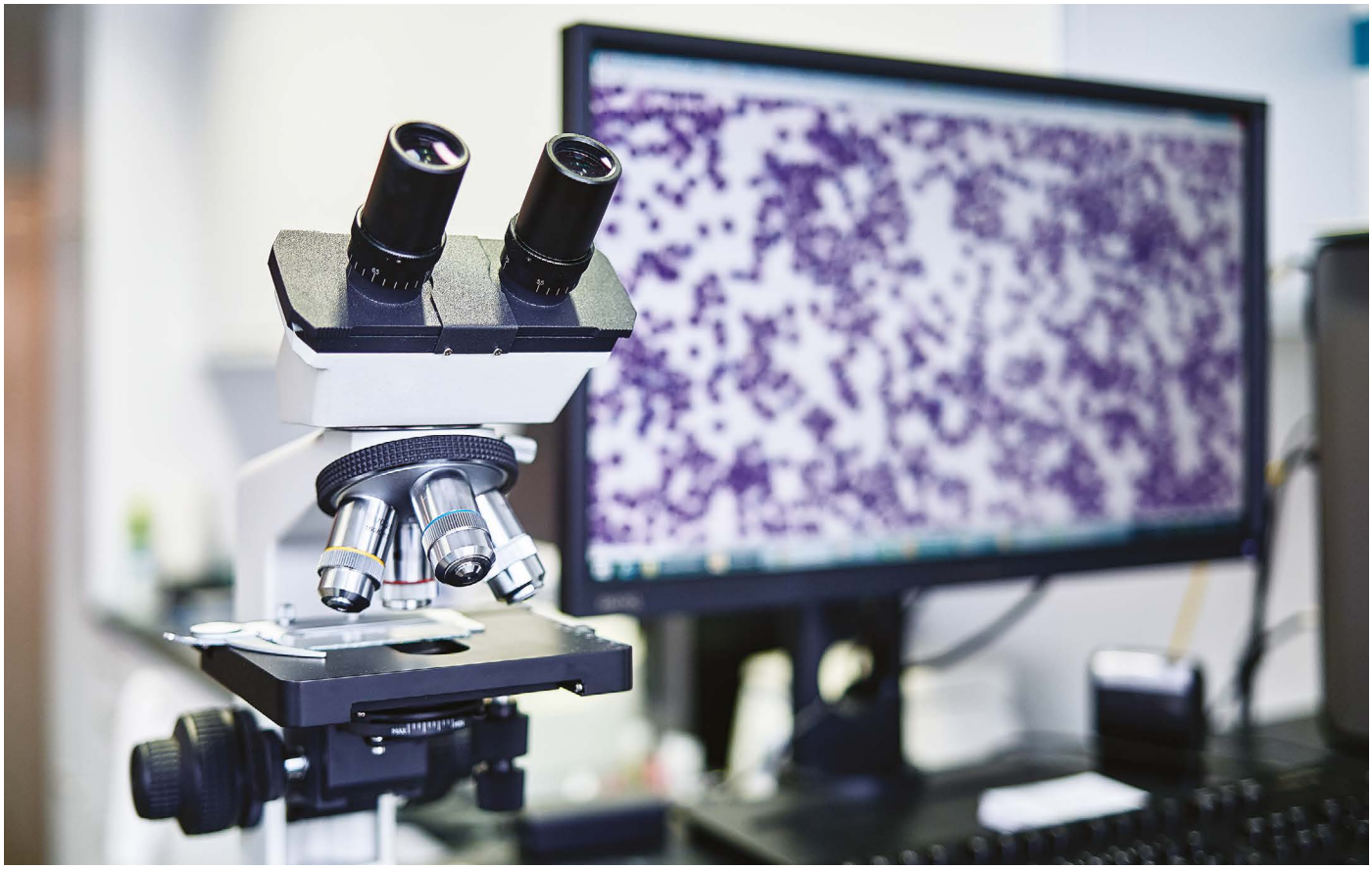

GETTY

The increasing digitization of microscope slides is making pathology more amenable to advances in AI.

HOW ARTIFICIAL INTELLIGENCE IS TRANSFORMING PATHOLOGY

Some researchers say that deep-learning 'foundation' models will revolutionize the field — but others are not so sure. **By Diana Kwon**

If you've ever had a biopsy, you — or at least, your excised tissues — have been seen by a pathologist. "Pathology is the cornerstone of diagnosis, especially when it comes to cancer," says Bo Wang, a computer scientist at the University of Toronto in Canada.

But pathologists are increasingly under strain. Globally, demand is outstripping supply, and many countries are facing shortages. At the same time, pathologists' jobs have become more demanding. They now involve not only more and more conventional tasks, such as sectioning and staining tissues, then viewing them under a microscope, but also tests that require extra tools and expertise, such as assays for genes and other molecular markers. For Wang and others, one solution to this growing problem could lie in artificial intelligence (AI).

AI tools can help pathologists in several ways: highlighting suspicious regions in the tissue, standardizing diagnoses and uncovering patterns invisible to the human eye, for instance. "They hold the potential to improve diagnostic accuracy, reproducibility and also efficiency," Wang says, "while also enabling new research directions for mining large-scale pathological and molecular data."

Over the past few decades, slides have increasingly been digitized, enabling pathologists to study samples on-screen rather than under a microscope — although many still prefer the microscope. The resulting images, which can encompass whole slides, have proved invaluable to computer scientists and biomedical engineers, who have used them to develop AI-based assistants. Moreover, the success of AI chatbots such as ChatGPT and DeepSeek has inspired researchers to apply similar techniques to pathology. "This is a very dynamic research area, with lots of new research coming up every day," Wang says. "It's very exciting."

Scientists have designed AI models to perform tasks such as classifying illnesses, predicting treatment outcomes and identifying biological markers of disease. Some have even produced chatbots that can assist physicians and researchers seeking to decipher the data hidden in stained slices of tissue. Such models "can essentially mimic the entire pathology process", from analysing slides and ordering tests to writing reports, says Faisal Mahmood, a computer scientist at Harvard Medical School in Boston, Massachusetts. "All of that is possible with technology today," he says.

But some researchers are wary. They say that AI models have not yet been validated sufficiently — and that the opaque nature of some models poses challenges to deploying them in the clinic. "At the end of the day, when these

tools go into the hospital, to the bedside of the patient, they need to provide reliable, accurate and robust results," says Hamid Tizhoosh, a computer scientist at Mayo Clinic in Rochester, Minnesota. "We are still waiting for those."

Building foundations

The earliest AI tools for pathology were designed to perform clearly defined tasks, such as detecting cancer in breast-tissue biopsy samples. But the advent of 'foundation' models — which can adapt to a broad range of applications that they haven't been specifically trained for — has provided an alternative approach.

Among the best-known foundation models are the large language models that drive generative-AI tools such as ChatGPT. However, ChatGPT was trained on much of the text on the Internet, and pathologists have no comparably vast resource with which to train their software. For Mahmood, a potential solution to this problem came in 2023, when researchers at tech giant Meta released DINOv2, a foundation model designed to perform visual tasks, such as image classification[1]. The study describing DINOv2 provided an important insight, Mahmood says — namely, that the diversity of a training data set was more important than its size.

By applying this principle, Mahmood and his team launched, in March 2024, what they describe as a general-purpose model for pathology, named UNI (ref. 2). They gathered a data set of more than 100 million images from 100,000 slides representing both diseased and healthy organs and tissues. The researchers then used the data set to train a self-supervised learning algorithm — a machine-learning model that teaches itself to detect patterns in large data sets. The team reported that UNI could outperform existing state-of-the-art computational-pathology models on dozens of classification tasks, including detecting cancer metastases and identifying various tumour subtypes in the breast and brain. The current version, UNI 2, has an expanded training data set, which includes more than 200 million images and 350,000 slides (see go.nature.com/3h5qkwb).

A second foundation model designed by the team used the same philosophy regarding diverse data sets, but also included images from pathology slides and text obtained from PubMed and other medical databases[3]. (Such models are called multimodal.)

Like UNI, the model — called CONCH (for Contrastive Learning from Captions for Histopathology) — could perform classification tasks, such as cancer subtyping, better than could other models, the researchers found. For example, it could distinguish between subtypes of cancer that contain mutations in the *BRCA* genes with more than 90% accuracy, whereas other models performed, for the most part, no better than would be expected by chance. It could also classify and caption images, retrieving text in response to image queries and vice versa, to produce graphics of the patterns seen in specific cancers. However, it was not as accurate in these tasks as it was for classification. In head-to-head evaluations, CONCH consistently outperformed baseline approaches even in cases where very few data points were available for downstream model training.

UNI and CONCH are publicly available on the model-sharing platform Hugging Face (see go.nature.com/44g24w2). Researchers have used them for a variety of applications, including grading and subtyping neural tumours called neuroblastomas, predicting treatment outcomes and identifying gene-expression biomarkers associated with specific diseases. With more than 1.5 million downloads and hundreds of citations, the models have "been used in ways that I never thought people would be using them for", Mahmood says. "I had no idea that so many people were interested in computational pathology."

Other groups have developed their own foundation models for pathology. Microsoft's GigaPath[4], for instance, is trained on more than 170,000 slides obtained from 28 US cancer centres to do tasks such as cancer subtyping. mSTAR (for Multimodal Self-taught Pretraining), designed by computer scientist Hao Chen at the Hong Kong University of Science and Technology and his team, folds together gene-expression profiles, images and text[5]. Also available on Hugging Face (see go.nature.com/3ylmauf), mSTAR was designed to detect metastases, to subtype cancers and to perform other tasks.

> **"The cleanest and best way to validate these models is through external, independent validation."**

Now, Mahmood and Chen's teams have built 'copilots' based on their models. In June 2024, Mahmood's team released PathChat — a generalist AI assistant that combined UNI with a large language model[6]. The resulting model was then fine-tuned with almost one million questions and answers using information derived from articles on PubMed, case reports and other sources. Pathologists can use it to have 'conversations' about uploaded images and generate reports, among other things. Licensed to Modella AI, a biomedical firm in Boston, Massachusetts, the chatbot received breakthrough-device designation from the US Food and Drug Administration earlier this year.

Similarly, Chen's team has developed SmartPath, a chatbot that Chen says is being tested in hospitals in China. Pathologists are going head to head with the tool in assessments of breast, lung and colon cancers.

Beyond classification tasks, both PathChat and SmartPath have been endowed with agent-like capabilities — the ability to plan, make decisions and act autonomously. According to Mahmood, this enables PathChat to streamline a pathologist's workflow — for example, by highlighting cases that are likely to be positive for a given disease, ordering further tests to support the diagnostic process and writing pathology reports.

Foundation models, says Jakob Kather, an oncologist at the Technical University of Dresden in Germany, represent "a really transformative technological advancement" in pathology — although they are yet to be approved by regulatory authorities. "I think it'll take about two or three years until these tools are widely available, clinical proven products," he says.

An AI revolution?

Not everyone is convinced that foundation models will bring about groundbreaking changes in medicine — at least, not in the short term.

One key concern is accuracy. Specifically, how to quantify it, says Anant Madabhushi, a biomedical engineer at Emory University in Atlanta, Georgia. Owing to a relative lack of data, most pathology-AI studies use a cross-validation approach, in which one chunk of a data set is reserved for training and another for testing. This can lead to issues such as overfitting, meaning that an algorithm performs well on data similar to the information that the model has encountered before but does poorly on disparate data.

"The problem with cross-validation is that it tends to provide fairly optimistic results," Madabhushi explains. "The cleanest and best way to validate these models is through external, independent validation, where the external test set is separate and distinct from the training set and ideally from a separate institution."

Furthermore, models also might not perform as well in the field as their developers suggest they do. In a study published in February[7], Tizhoosh and his colleagues put a handful of pathology foundation models to the test, including UNI and GigaPath. Using a zero-shot approach, in which a model is tested on a data set it hasn't yet encountered (in this case, data from the Cancer Genome Atlas, which contains around 11,000 slides from more than 9,000 individuals), the team found that the assessed models were, on average, less accurate at identifying cancers than a coin toss would be — although some models did perform better for specific organs, such as the kidneys.

According to Tizhoosh, the discrepancy between published performance and what his team observed could come down to

"fine-tuning". Researchers typically tweak models before use by providing numerous examples related to their specific question, but Tizhoosh's team used the models as is. Still, these findings suggest that AI-based pathology tools might not be as revolutionary as their designers state, he says. "I'm worried that they are overpromising, and that this will create a new wave of disappointment in AI."

Several groups have launched efforts to standardize validation and benchmarking processes. For example, Tizhoosh, along with colleagues at the Memorial Sloan Kettering Cancer Center in New York City and the University of Texas MD Anderson Cancer Center in Houston, is preparing a challenge in which participants will be given 150 million images on which to train their models. They will then submit the models for independent testing. "We are hoping that from that undertaking, which will be closed by the end of the year, a set of rules and guidelines could emerge," Tizhoosh says.

Another group, led by computer scientist Francesco Ciompi at the Radboud University Medical Center in Nijmegen, the Netherlands, has also launched several such challenges. One, called UNICORN (Unified Benchmark for Imaging in Computational Pathology, Radiology and Natural Language) will test multimodal foundation models on a series of tasks including pathology and radiology. "The goal is to see how well these foundation models do without much fine-tuning," Ciompi says.

No simple task

Even those enthusiastic about foundation models acknowledge that validation is no simple task. The models are designed to be open ended and adaptable. The "most conservative" way to assess them, says Kather, is to test every application. "So, if you have 1,000 different use cases, you have to collect hundreds of tissue slides for every single one of these use cases and apply your model to that."

Kather says that there are ongoing discussions around radically different approaches for measuring performance. For instance, he suggests, as AI models become more human-like in their abilities, perhaps they should be tested the same way people are. "You don't evaluate a human in every single use case, you evaluate their general understanding of things – you pick some examples and evaluate their performance."

There are other issues, too, including generalizability: making sure that these tools work across a diverse range of people. In 2021, for instance, Oncotype DX, a molecular test that assesses whether people with breast cancer could benefit from chemotherapy, came under fire. Researchers found that, despite having been on the market for at least two decades, the test was much less effective for Black women than for white women[8]. "If you're not intentional about how you're developing and also validating these algorithms, you're going to run into these catastrophic errors," Madabhushi says.

There's also the problem of hallucination, in which chatbots fabricate responses to queries. In medicine, an incorrect answer could lead to a wrong or missed diagnosis. "How do you measure the safety and the interpretability of these models to alleviate risks when it comes to patient diagnosis?" Wang says. "Regulators like the FDA simply do not have any guidelines for generative models in the health-care space."

And then there's the fact that foundation models are effectively black boxes, meaning that it can be difficult to work out how they arrived at their answers. "Foundation models are exciting, but we still lack an understanding of what these models are picking up," says Madabhushi.

Madabhushi works on what he calls "explainable AI" – models based on conventional techniques in which researchers program algorithms to pinpoint specific biological features associated with diseases. For example, his team has developed models that search for specific patterns of collagen fibres that identify early-stage breast cancer[9] and arrangements of immune cells that predict outcomes in individuals with cancer who receive immunotherapy[10]. (Madabhushi co-founded Picture Health, a biotechnology company in Cleveland, Ohio, that has licensed these technologies and is working towards regulatory approval.)

Other researchers are working on opening the models' black boxes – at least to some extent. Chen, for one, says that he and his team are working on ways to trace the steps their models take to get to the answers, in hopes of shedding light on how these algorithms make the decisions they do. "We want our models to be accurate and trustworthy," Chen says. "And for doctors, one of the most important things is explainability."

The field has a long way to go, but Chen is optimistic. "This is just the beginning," he says. "Some people may overestimate the power of this technology – but it can also easily be underestimated in the long term."

Diana Kwon is a freelance science journalist based in Berlin.

1. Oquab, M. *et al.* Preprint at arXiv https://doi.org/10.48550/arXiv.2304.07193 (2023).
2. Chen, R. J. *et al. Nature Med.* **30**, 850–862 (2024).
3. Lu, M. Y. *et al. Nature Med.* **30**, 863–874 (2024).
4. Xu, H. *et al. Nature* **630**, 181–188 (2024).
5. Xu, Y. *et al.* Preprint at arXiv https://doi.org/10.48550/arXiv.2407.15362 (2024).
6. Lu, M. Y. *et al. Nature* **634**, 466–473 (2024).
7. Alfasly, S. *et al. Sci. Rep.* **15**, 3990 (2025).
8. Hoskins, K. F., Danciu, O. C., Ko, N. Y. & Calip, G. S. *JAMA Oncol.* **7**, 370–378 (2021).
9. Li, H. *et al. npj Breast Cancer* **7**, 104 (2021).
10. Wang, X. *et al. Sci. Adv.* **8**, eabn3966 (2022).

Correction

The 20 tasks that UNICORN will be testing are not exclusively pathology-related as originally stated, but cover a range of things including pathology and radiology.