The ComPatAI consortium uses large datasets to create an AI learning model for pathology

Pekka Ruusuvuori, Associate Professor of the Institute of Biomedicine at the University of Turku, leads the ComPatAI consortium, which is developing new ways to model histopathological tissue samples with generative and predictive AI.

In medicine, histological samples are used to assess a patient's need for treatment. The consortium's goal is to use big data to create AI models that would produce more accurate diagnostic information in pathology.

In addition, they are developing virtual histological staining models based on generative AI. Besides Ruusuvuori, the consortium consists of research director and Adjunct Professor **Leena Latonen** of University of Eastern Finland and **Teemu Tolonen**, Adjunct Professor and chief physician at the department of pathology at the Fimlab laboratories.

The ComPatAI consortium focuses on analysing histological samples related to breast and prostate cancer. Using digitalised images allows the researchers to measure and automatically compute different cell types.

"Our work has focused mainly on prostate and breast cancer. There is ample data available on these types of cancers, as they are the most commonly encountered cancers in men and women. However, we want to create a very general-purpose model that could then be further refined for different and new applications."

According to Ruusuvuori, the field of pathology is becoming increasingly digitalised. He says that in this sense, the Finnish pathological community is among the pioneers.

"In Tampere and Turku, we have moved completely to using digital pathology in diagnostics. Each time a sample is taken, it is scanned into a high-resolution digital image. There is a lot of routine diagnostics. As the population ages, we encounter more and more patients with cancer. This also means that there are loads of data coming in."

600,000 histological whole-slide images

The consortium receives scanned images of histological slides from Fimlab, the largest healthcare laboratory company in Finland. Fimlab's clientele includes hospitals, health centres, occupational healthcare service providers and private medical stations. The Finnish Medicines Agency FIMEA has currently granted the consortium a licence for using data from 160,050 cases, which translates to a total of approximately 600,000 slide images. Together, the images add up to about 0.8 petabytes of data, meaning that each file accounts for approximately 1.3 gigabytes. The massive amounts of data are currently being anonymised and transferred to the LUMI supercomputer in the CSC – IT Center for Science, an ELIXIR node in Finland. The project is one of the largest data transfers made to LUMI so far.

"It is incredible that we get to use these data for our research. We want to use this data to create AI solutions that function well in pathological work", Ruusuvuori explains.

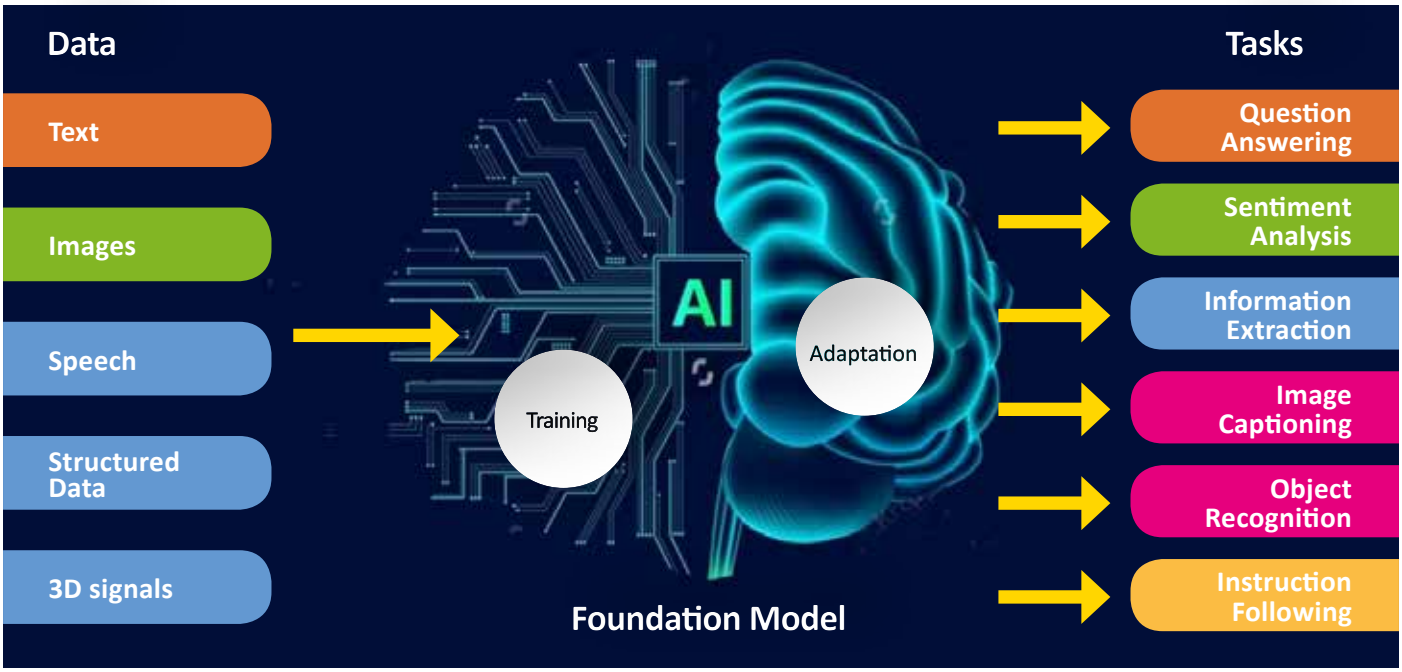
The researchers are currently set to have up to 2.5 million digitised whole-slide images at their disposal by the end of the project. This corresponds to a total of three petabytes of data.

"We have been granted permission to use technically all data produced in Fimlab's routine digital pathology operations."

Neural network learns from the image things that the human eye cannot perceive

Pekka Ruusuvuori has a strong background in signal processing, and he specialises in image analysis. Ruusuvuori is interested in how deep neural networks that are used in AI applications could be developed to be a better fit for diverse use cases.

According to him, machine can generally be taught to recognise the same things that humans would pick up on. It can learn to tell apart different tissue types or to distinguish



: Overview of the foundation model. Foundation models rely on deep neural networks and their ability to learn complex patterns and structures based on different datasets. They are increasingly used for analysing image data. The models learn to combine visual cues such as colours, shapes or textures with semantic information, like the meaning or object of certain images. They break the images down into pixel-level information to learn about more complex features. A mathematical technique called self-attention helps them to prioritise the correct elements in the picture and understand how they relate to one another.

cancerous tissues from healthy ones. It can be used to measure different factors in images or cells, such as how aggressive a cancer is and how far it has progressed. Artificial intelligence can identify cancerous areas in tissue samples before examination by a pathologist. It may also suggest a score based on the data it has assessed. For example, prostate cancer tumours are given a Gleason score, which indicates how aggressive or advanced the disease is.

"It's entirely possible to train AI to perform many tasks that human pathologists usually take care of."

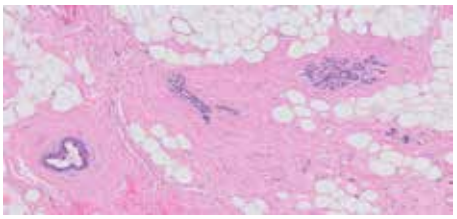
"Previously machine learning models have been built with a certain variable and teaching material that shows a certain object appearing in a specific part of the image and which score this finding corresponds to. It would take countless hours of work for us to mark this information on all the hundreds of thousands of images we are using."

This annotation data has previously played a key role in teaching artificial intelligence to automatically detect abnormalities such as cancer cells in the samples. Ruusuvuori says that algorithms have been improved and are consequently able to use unannotated raw data.

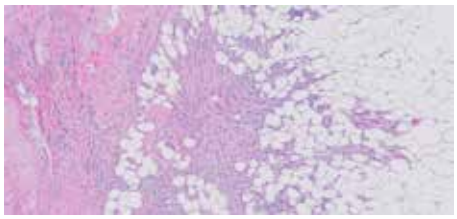
"I think the most interesting thing about what we are doing with these algorithms is what else we can extract from these images. In other words, to look at the features that machines can detect but humans cannot. These slide images include all visual data that we have. If there is a statistical link to be found there, the machine learning algorithm will find it. However, these links may be extremely complex. Modern neural networks can accurately detect complex links between spatial data and the predicted variable. These things can be very difficult for us humans to grasp."

Together with his research group, Ruusuvuori has been able to successfully predict gene expression and mutations directly from histological images. Gene expression refers to the process of a cell producing the molecules that is encoded in its DNA. The gene expression varies across different tissues. Based on the images, AI can detect miniscule changes that are invisible to the human eye.

"Based on the images, the machine can identify effects of gene expression in cells and tissues. It can detect even the slightest phenotype variations, including those that we as humans are not trained to see. I want to highlight that so far, we only have indicative results, and that the method will not work for all tissue types or genes. Some gene expressions



Healthy Tissue: Within the connective tissue, regular milk ducts and lobules (glandular segments) are observed. Image: Fimlab



Breast Cancer: The tumor forms fibrous structures and small clusters that grow irregularly into the stroma (the connective tissue network of the tumor) and fat. Image: Fimlab

Pekka Ruusuvuori's research project (Towards an AI-enabled Computational Pathology) has received funding from the Research Council of Finland. It is also a part of the LUMI Extreme Scale Access project, which tests how high-performance computing can make use of public data. Ruusuvuori and Leena Latonen are also working on another project funded by the Research Council of Finland. This project focuses on high-performance computing and virtual staining of histological samples. Funding from the Research Council helps to increase the use of European High-Performance Computing (EuroHPC) resources and the LUMI supercomputer for scientific research on flagship topics.

do not lead to tissue-level changes that could be predicted from a whole slide image."

ComPatAI consortium is currently developing a so-called foundation model for utilising large datasets. As the name suggests, this model would create a general-purpose foundation for developing further AI solutions. The model is trained in histology based on a large set of samples, without target variables or annotation data.

"When you start teaching a model like this to recognise diseases such as breast or prostate cancer, it starts to learn to perform the task it has been given. This will allow us to reach more accurate solutions much faster than before. It allows us to use large, unannotated datasets. This is a big step forward for us."

ComPatAI consortium is currently building its own foundation model based on a Finnish dataset.

"This is basic research that will allow us to be among the first to further refine these models in Finland. I do not want us to rely solely on big foreign firms or research groups, and instead wish that we would be able to be build a model based on Finnish data. We have high-quality, population-level cohort data that need to put to good use. I hope that this will lead to the establishment of companies in Finland that will develop solutions to benefit patients in routine diagnostics."

One of the key questions is how quickly data can be transferred and used. Computing and data storage capacity are constantly in high demand. This is where the services provided by CSC, the Finnish ELIXIR node, come in.

"We're extremely pleased with the support CSC has given us, since this is an exceptionally large project that uses very large datasets. We are in a privileged position because we have the support of an organisation like CSC. This is a clear competitive advantage for us, and we really appreciate it."

Digital pathological data and other potentially sensitive health data types, such as registry and omics datasets, are going to become more readily available through CSC's data secure user environment.

"We're just getting started with the development work", says Tommi Nyrönen, who leads the ELIXIR Finland Node.

"ELIXIR's node in Finland has helped transform the biomedical resources required by ComPatAI to a platform service operated by CSC. The CSC Sensitive Data platform emerged from this need. However, it continues to serve various other researcher projects in the field. One of these is the EU's digital pathology archive initiative bigpicture.eu, which is set to launch in 2026. It is a sustainable solution for managing digital pathology datasets and bringing them to high-performance computing services across Europe."

14.12.2024 | ***Ari Turunen***

MORE INFORMATION

Ruusuvuorilab
https://ruusuvuorilab.utu.fi

Fimlab
https://fimlab.fi/en/

University of Turku
https://www.utu.fi/en

CSC – IT Center for Science
is a non-profit, state-owned company administered by the Ministry of Education and Culture. CSC maintains and develops the state-owned, centralised IT infrastructure.
http://www.csc.fi
https://research.csc.fi/cloud-computing

ELIXIR
builds infrastructure in support of the biological sector. It brings together the leading organisations of 21 European countries and the EMBL European Molecular Biology Laboratory to form a common infrastructure for biological information. CSC – IT Center for Science is the Finnish centre within this infrastructure.
http://www.elixir-finland.org
http://www.elixir-europe.org

ELIXIR FINLAND
Tel. +358 9 457 2821s ● e-mail: servicedesk@csc.fi
www.elixir-europe.org/about-us/who-we-are/nodes/finland

www.elixir-finland.org

ELIXIR HEAD OFFICE
EMBL-European Bioinformatics Institute
www.elixir-europe.org