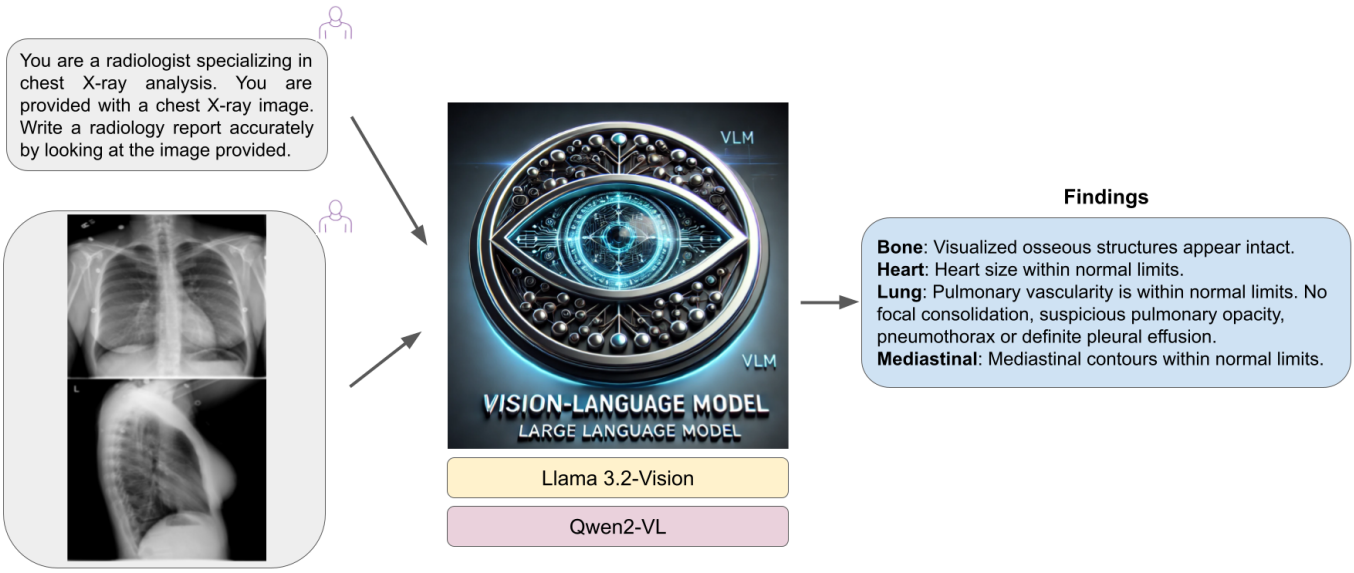
Portfolio in AI/ML/Computer Vision

Hello, guys! I'm a Scientist in Machine Learning and Computational Science. Over the past 5 years, I've specialized in creating machine-learning products for the biotech and oil and gas industries. I love the intersection between science and real-world problems. I typically work on private repositories, but here are a few standout projects we've published. Email: 77777aidan@gmail.com

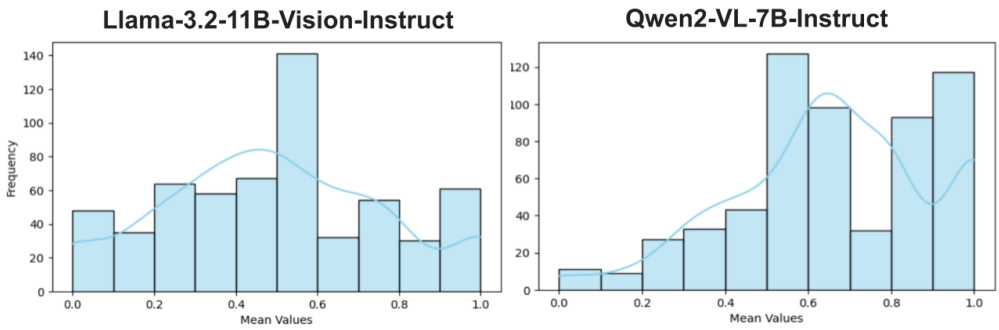
Multi-Modal Vision-Language Transformers for Automated Radiology Report Generation from X-ray Images

I trained (fine-tuned) a multi-modal LLM model to predict the findings on X-ray images to help radiologists in their day-to-day work. The present global shortage of radiologists limits access to specialist care and imposes heavy workloads on radiologists, leading to unwanted delays and errors in clinical decisions. This model can automate X-ray report generation, enhance diagnostic accuracy, and facilitate clinicians in providing timely and effective patient care.

Unfortunately, generating radiology reports remains an unsolved challenge. Here, I present an approach to tackle this challenge and showcase the use of Vision-Language LLM to predict radiology reports. The model inputs multi-modal image and text data and provides an X-ray report. I use the Llama 3.2-Vision and Qwen2-VL models and fine-tune them on the 2900+ chest X-ray images from the public IU X-Ray dataset.



For the model performance evaluation of the clinical quality of artificial intelligence (AI)-generated reports, I use the GREEN (Generative Radiology Report Evaluation and Error Notation) metric. It utilizes language models to identify and explain clinically significant errors in radiology reports. Other evaluation metrics can also be used, but they either fail to consider factual correctness, such as BLEU and ROUGE or have limited interpretability, like F1CheXpert and F1RadGraph (Ostmeier et al., 2024). I get the following distribution of GREEN scores evaluated on the per report sample image:



GREEN uses advanced language models to:

- Compare a machine-generated radiology report to a trusted, expert-written reference report.
- Spot clinically important mistakes-for example, if the AI says “no pneumonia” when there actually is pneumonia.
- Explain the errors in plain language, so users know exactly what went wrong and why.

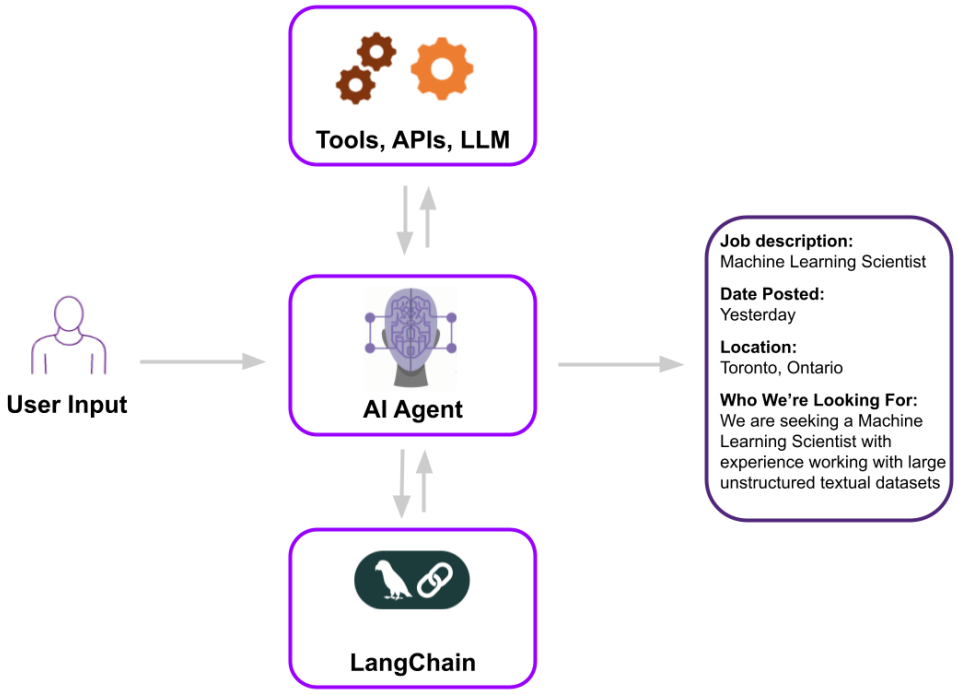
AI agent for job search using LangChain

An AI agent is a specialized software solution built on top of large language models (LLMs) and fine-tuned to automate specific, critical tasks across various industries. Examples include:

- AI tax accountant that manages tax preparation and advisory services,
- AI medical biller that handles patient records and submits claims,
- AI phone support agent that responds to customer inquiries in real-time,
- AI compliance agent that ensures regulatory adherence, and an AI quality assurance tester that automates software testing processes.

These agents enhance efficiency by focusing on niche areas, and providing accurate and effective solutions. Here, I have developed an AI agent capable of utilizing multiple tools for efficient job searching and company research. The agent leverages LangChain, OpenAI’s chat models, and external tools such as Google Jobs and Wikipedia search to retrieve structured information about job openings. A LangChain agent consists of various components, including chat LLM models, prompt templates, external tools, and other integrations.

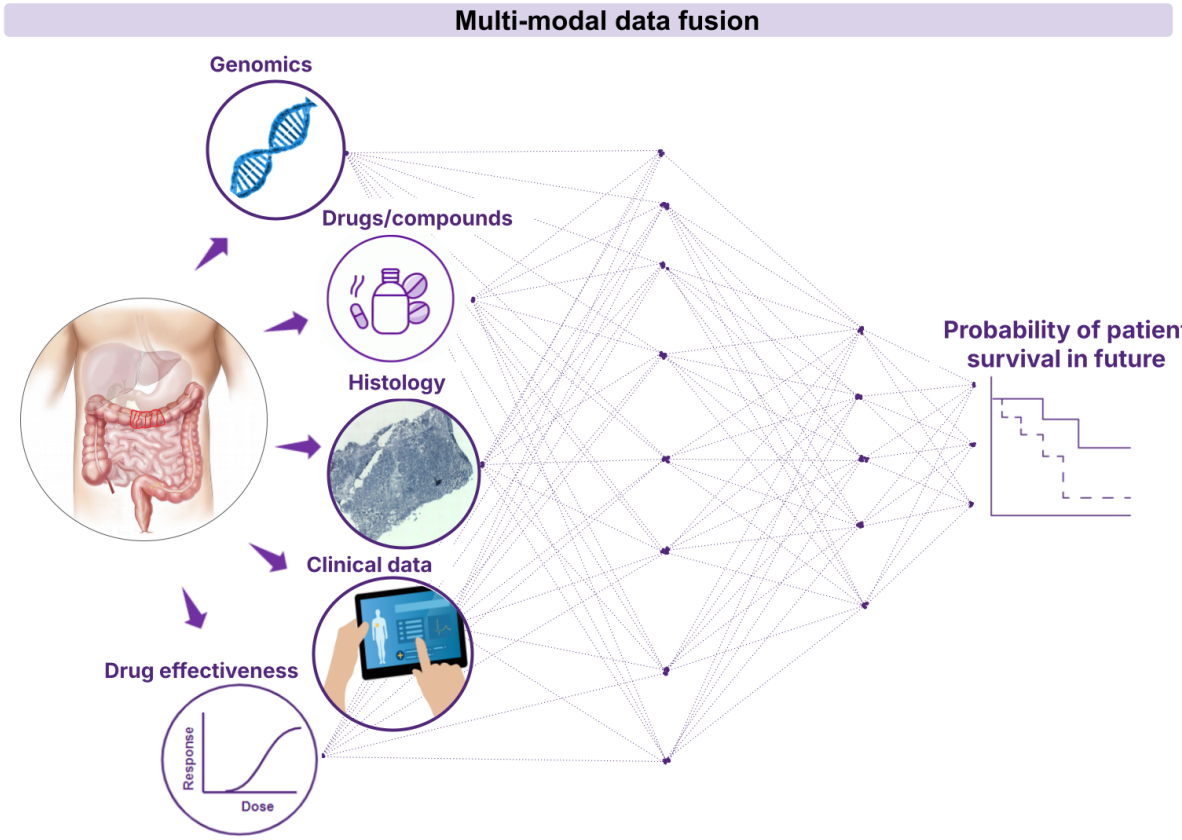
I have implemented a ReAct agent (reason and act), which offers a more structured approach to building AI systems. The agent is provided with specific instructions, such as acting as a recruiter or an assistant, and generates responses based on user queries. By interacting with real-world data sources, it ensures dynamic and real-time functionality. The results demonstrate how the agent selects the appropriate tool—such as Wikipedia for answering company-related queries or Google Jobs for job searches—based on the prompt given to the LLM chat model.



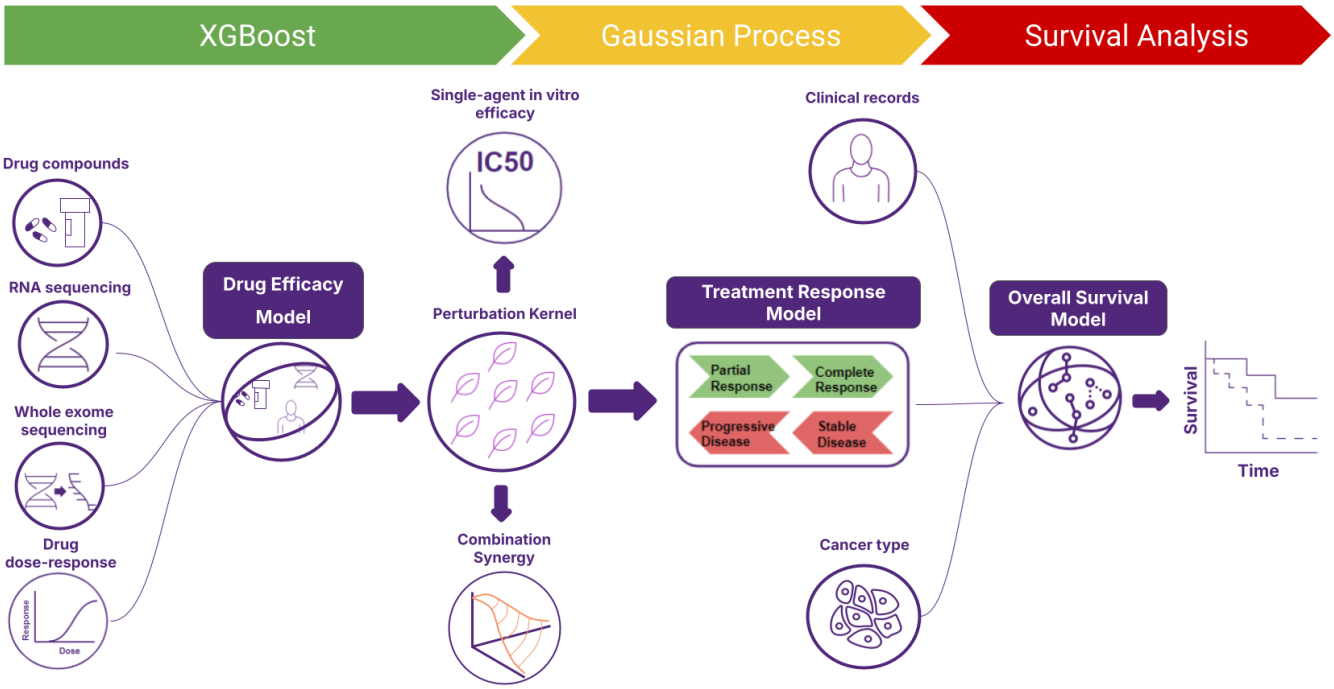
Integrated ML Predictor of Clinical Trials for Drug Discovery

We have developed a method that employs Bayesian statistics to accurately forecast the outcomes of clinical trials in the course of novel drug development. The development of an oncology drug currently incurs a cost exceeding £4 billion, given the high failure rate of approximately 95%. Our proposed Digital Twin can simulate a clinical trial and predict novel drug outcomes, thereby improving and mitigating risks in the clinical development of oncology therapeutics.

- It uses multi-modal data: genetics (like RNAseq), clinical, image data and chemical compounds data
- We integrate XGBoost, Gaussian Process and Survival Modelling into one model
- We validate our model against past clinical trials and use standard ML validation methods



We validate our model by comparing its outputs against the actual historical clinical trials blindly and un-blindly. We simulated digital twin trial arms for single chemotherapy drugs and combinations to predict treatment response. TCGA data was used as input, and treatment predictions were compared to the results of eight published historical phase 2 and phase 3 clinical studies (1997 - 2018), as these trials were assessable with the input data available. We compared the predicted log odds ratio (OR) generated by the digital twin model for Overall Response Rates (ORR) for each treatment arm tested in the clinical study, and then compared this against the reported log odds ratios (log OR) from the actual trial. We started with single-agent predictions, then progressively increased the complexity through combinations and heterogeneous treatments in more sophisticated clinical trial designs.

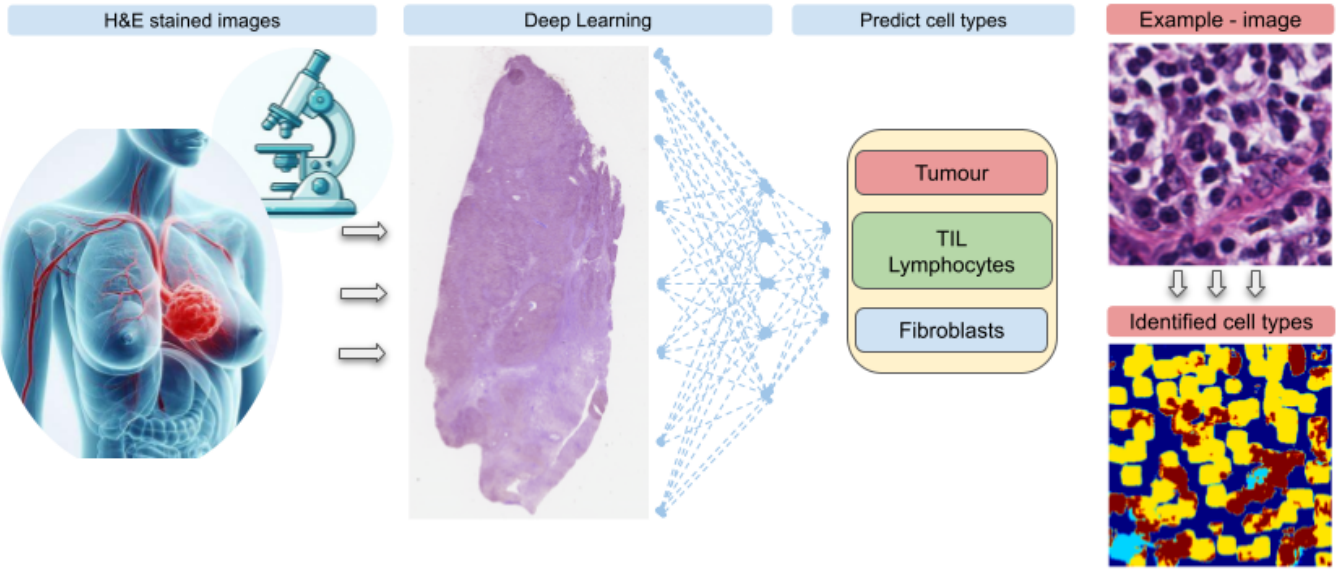


The paper submitted to NEJM AI. I have an older preprint on the MedRxiv, the newer and significantly better is being reviewed. Let me know if you'd like to receive the newer version.

Computer Vision ML to Identify Cell Types on the Medical Images

We developed a deep-learning model to identify types of cells (tumour cells, lymphocytes, and fibroblasts) from medical images. The model was utilized to investigate the impact of image features on the modelling of clinical trials for drug discovery and predicting treatment outcomes. Additionally, it could assist pathologists, who spend a considerable amount of time on diagnostics, by improving their productivity.

The model was trained with annotated images from a breast cancer dataset and validated using standard machine learning validations and by another expert pathologist. The results demonstrated that the model could accurately identify the cells, especially tumour-infiltrating lymphocytes (TILs), which it identified even more accurately than a pathologist.

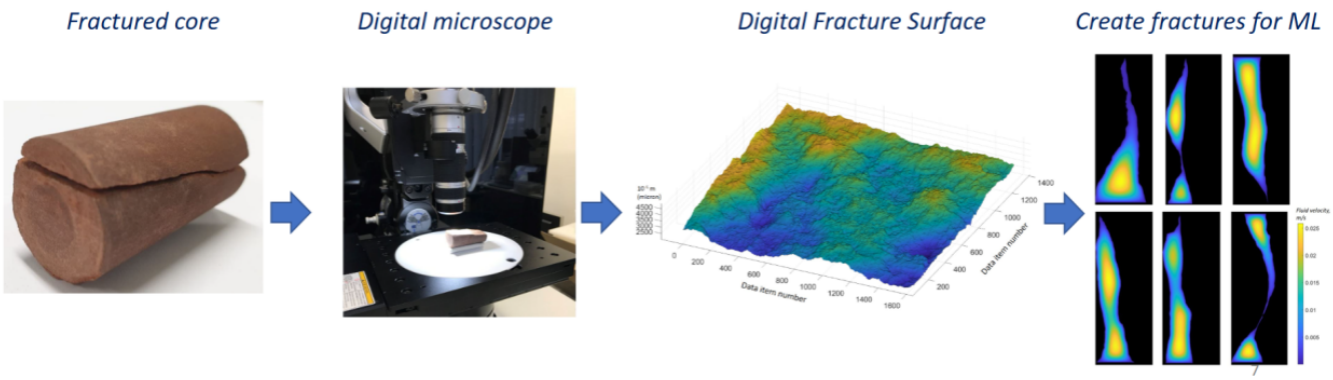


We employed a U-Net architecture for semantic segmentation, training the model on the NuCLS dataset annotated for cell types. The dataset, comprising images from the TCGA for breast cancer, was split into training and validation/testing sets. The model's performance demonstrates promising results with an AUROC of 0.864 and 0.901, along with balanced and standard accuracies. The model offers a tool that can enhance precision treatment by integrating it into complex predictive modelling systems.

We presented it at the AACR conference in 2023, the poster is available to download here.

Hybrid “Deep Learning + Physics” Computer Vision Model

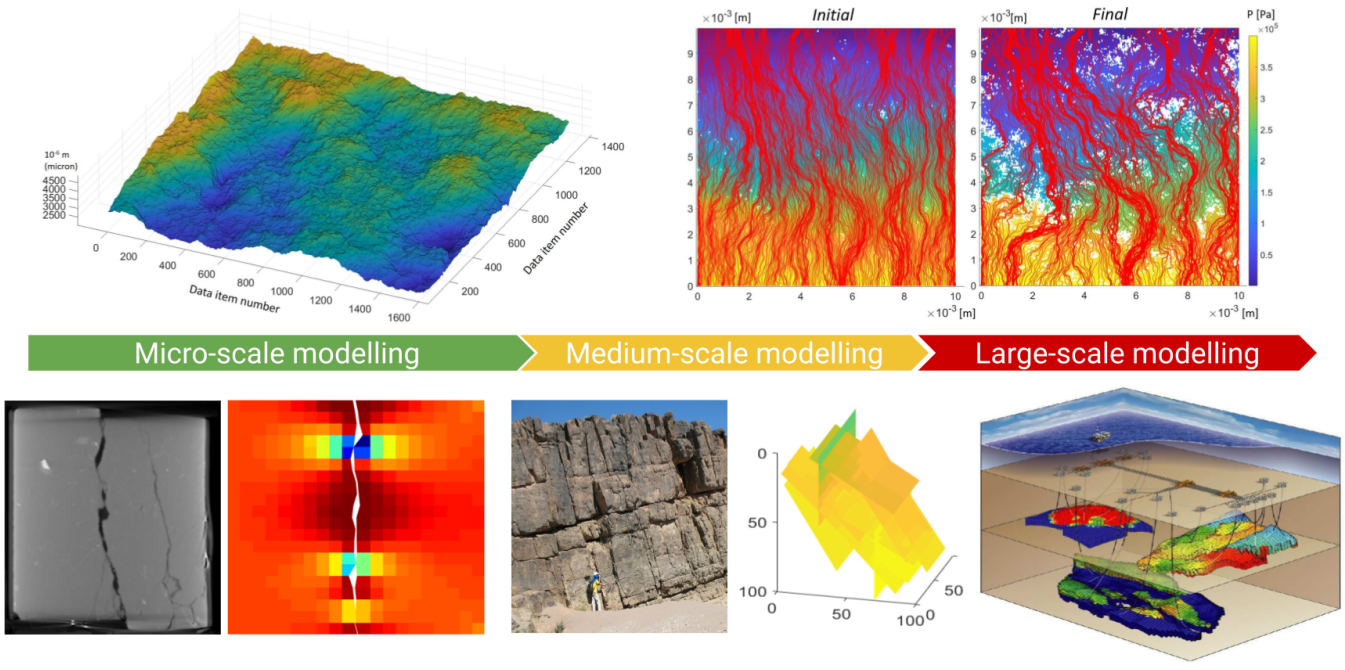
Here, I designed and created a deep learning model that combines computer vision deep learning and physics for the physics-based simulation. Solving multi-physics problems usually requires expensive, high-performance computers and complex code. The model helps and explores how machine learning can reduce computation time in these kinds of problems. It uses a deep learning model, specifically a convolutional neural network, to predict rough fracture permeability from digital images during the fracture deformation process. Even in extrapolation tests with different fracture roughness, the model maintained high accuracy with about 8% MAPE. I show that my method is able to speed up the numerical simulation up to 20 times faster than the conventional fully physics-based methods.



I published it in the Engineering Applications of Artificial Intelligence journal in 2023.

Physics-based Computer Vision Method for Energy, Oil&Gas, Climate Change and Earthquakes projects

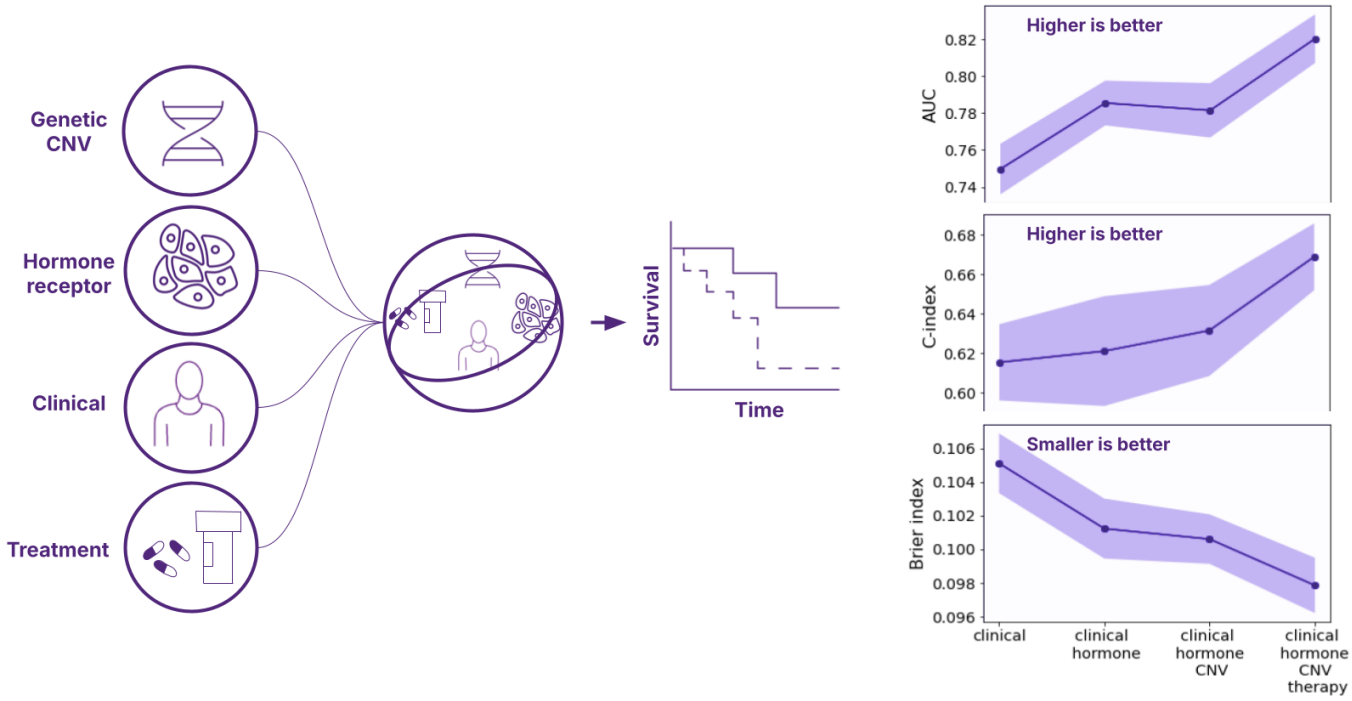
I created a research software tool (Physics-based Computer Vision Method) and methodology that takes digital images as inputs, performs predictive modelling, and identifies key performance indicators (KPIs) for energy, oil&gas and climate change projects. The model focuses on understanding and simulating how stress affects the permeability of rough fracture surfaces during the surface deformation process, using a combination of numerical contact mechanics, numerical modelling and the Stokes equation. This approach allows for the simulation of mechanical deformation and fluid flow in natural fractures with complex geometries. The software accurately predicts the stress-permeability relationship, helping to provide valuable insights for hydro-mechanical studies of geological formations. This tool significantly reduces computation time, providing quick and accurate results that can inform better decision-making in energy, oil&gas and climate change projects, such as GCCS.



I published it in the Transport in Porous Media, which focuses on the research on the physical and chemical aspects of the transport of mass of a fluid phase, the mass of a component of a phase, momentum and energy, in single and multiphase flow in the porous medium domain.

Multi-modal data-based ML for cancer survival

Accurate modelling of the impact of patient-specific features and cancer treatments on survival allows the assignment of targeted therapy. Delivering personalized medicine to select the “best” cancer treatment for the individual is challenging. There is a need to build a multi-source data-driven model for the survival analysis of breast cancer. I developed an ML model that integrates multi-modal data - genetic, hormone, clinical, and therapy data - to predict survival for breast cancer patients.



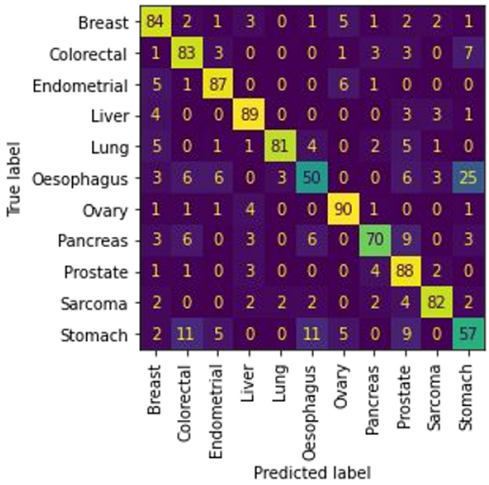
I show that combining multi-modal data features enhances model predictive accuracy, up to AUC 0.82 on unseen test data. Results depict differential accuracy measured by a time-integrated Area Under the Curve (AUC), weighted Concordance Index and Brier index. The predictive accuracy improved stepwise by adding additional relevant data types.

This model is part of the Integrated ML Predictor of Clinical Trials and answers questions like "What is the recommended treatment for a 52-year-old patient with stage 3 breast cancer, ER+ve, HER2-ve, genetic information e.g FGFR2 copy number gain?"

I presented it orally at the AACR conference in 2023, the poster is available to download here.

ML to predict tissue of origin from mutation data

Cancer of Unknown Primary (CUP) is a clinical condition with a poor prognosis. Patients present with metastatic tumours for which the primary tissue of origin cannot be easily determined. The current standard of care relies on identifying the primary tissue of origin using radiological investigations, tumour marker assessment, and biopsy. For one-third of CUP patients, the origin of the primary tumour cannot be found, making it difficult to choose the best chemotherapy regime and preventing access to targeted and immune therapies.



	Top 1 acc	Top 2 acc	Top 3 acc	Precision	Recall	F1 score	Training size	Test size
Endometrial	0.90	0.95	0.98	0.86	0.87	0.87	400	100
Breast	0.88	0.97	0.98	0.88	0.84	0.86	756	189
Prostate	0.88	0.95	0.98	0.75	0.88	0.81	380	95
Lung	0.87	0.91	0.94	0.98	0.81	0.88	396	99
Ovary	0.87	0.94	0.97	0.77	0.90	0.83	312	78
Colorectal	0.85	0.97	0.98	0.83	0.83	0.83	460	115
Liver	0.83	0.92	0.96	0.82	0.89	0.85	288	72
Sarcoma	0.78	0.87	0.98	0.80	0.82	0.81	180	45
Pancreas	0.73	0.79	0.85	0.62	0.70	0.66	132	33
Stomach	0.71	0.91	0.94	0.68	0.58	0.62	340	85
Oesophagus	0.50	0.75	0.83	0.51	0.50	0.51	144	36
mean	**0.80**	**0.90**	**0.94**	**0.77**	**0.78**	**0.78**	**344**	**86**

We train a Machine Learning classifier to predict the tissue type of a solid tumour using only whole exome somatic mutation information.

We presented this study at the AACR conference, where I am the 3rd author.