Bocconi Student for Machine Learning - BSML

Digital Twin for Tumor Evolution and Metastasis Prediction

Motivation

Cancer remains one of the most complex and heterogeneous diseases. Advances in data science, machine learning, and artificial intelligence are transforming this landscape, enabling models that learn from massive clinical and molecular datasets. These technologies allow physicians to detect patterns invisible to the human eye and to predict how a tumor may evolve or respond to treatment. Digital twins embody this revolution: a virtual, AI-driven replica of a patient's tumor that can simulate biological behavior and forecast future changes.

What Is a Digital Twin?

A **digital twin** is a virtual, data-driven replica of a physical system, powered by machine learning, data science, and artificial intelligence. In oncology, it represents the biological and clinical behavior of a tumor in silico, continuously learning and updating as new patient data arrive. The integration of ML models allows the twin to recognize patterns, refine simulations, and enhance predictive accuracy, enabling simulation of disease progression, therapy response, and "what if" scenarios.

Our Goal

To develop a patient-specific digital twin that:

- Simulates the disease evolution dynamically over time.
- Predicts the risk of metastasis and treatment resistance.
- Integrates biological, clinical, and imaging data for realistic modeling.

Project Structure

This project focuses specifically on **Non-Small Cell Lung Cancer (NSCLC)**, the most prevalent type of lung cancer and one of the most suitable cases for digital twin development.

NSCLC provides abundant biological and clinical data, making it ideal for studying tumor evolution and testing predictive models.

The project is divided into two main phases (two semesters), each producing a concrete outcome.

- **Semester 1: Simulation Core (Biological Foundation)**

  Objective: Build a biologically grounded simulation model for NSCLC tumor growth. At this stage, the team defines all **biomedical variables** that describe tumor behavior, such as cell proliferation rate, apoptosis probability, nutrient diffusion, oxygen consumption, and mutation frequency.
  These variables are used to establish **biological priors**, which represent the initial probabilistic values or expected ranges based on literature and experimental data.
  Using these priors, the team will generate **synthetic datasets:** simulated data that imitate realistic tumor behaviors under different biological conditions. This allows testing and refining the model before introducing real clinical data.

  Expected outcome: a working NSCLC tumor simulator that visually reproduces tumor dynamics (growth, resource diffusion, mutation effects) and produces datasets consistent with biological expectations.
  To achieve this, **dependencies and distributions among the biological variables** will be analyzed to gain a clear understanding of the mechanisms defining NSCLC formation and growth. The goal of the first semester is to build a simulation as complete and realistic as possible, conditioned on the selected biological variables and their interactions.

- **Semester 2: Predictive Twin**
  The synthetic data from the first semester will serve as a bridge to calibrate and test the model. Then, the team will integrate **real datasets** from sources like TCGA, Human Cell Atlas, and TCIA to tune the parameters and validate the predictive accuracy. Machine learning techniques—such as LSTM, GNNs, and survival analysis models—will be applied to enable prediction of tumor progression, metastasis, and therapy response.

  Expected outcome: a predictive digital twin of NSCLC capable of simulating and forecasting tumor evolution, presented through an **interactive dashboard** that visualizes both simulated and real patient data.

Profiles and Skills We Are Looking For

We are looking for collaborators with backgrounds that bridge medicine, biomedical sciences, and data science. Given the interdisciplinary nature of the project, profiles from the Master of Science in Data Analytics and Artificial Intelligence in Health Sciences are especially relevant.

The following profiles would be particularly valuable to the project:
- Medical or biomedical knowledge with an interest in computational approaches
- Data science and machine learning
- Programming (Python, R) for data modeling or simulation

- Communication skills: capacity to explain technical concepts clearly to non-technical peers and understand clinical perspectives.

Contact Information

Bocconi Students for Machine Learning (BSML)
A BSML Project — All rights reserved
Project Lead: Roberta Claps
Email: roberta.claps@studbocconi.it
Phone: +39 3469676330

Contact us if interested or for any inquiry.