INTERNSHIP PROPOSAL

Master 2 or Engineer (COMPUTER SCIENCE or BIONFORMATICS)

Title: Counterfactual interpretation of deep model for biomarker discovery

Keywords: Neural networks, Counterfactual interpretation, XAI, gene expression.

Supervisors:

- Farida Zehraoui, AROB@S Team, IBISC laboratory, Paris-Saclay University, Univ. Evry
- Blaise Hanczar, AROB@S Team, IBISC laboratory, Paris-Saclay University, Univ. Evry

Contact: farida.zehraoui@univ-evry.fr

Duration of internship: 6 months

Location: IBISC Laboratory, IBGBI, University of Evry, 23 Boulevard de France, 91000 Evr

Description:

Deep learning is expected to have a pivotal role in diagnostic and therapeutic decision-making. Models trained on genomic data enable the prediction of diverse patient phenotypes with remarkable accuracy. Therefore, comprehending the key factors underpinning the decisions of the models becomes crucial [1] . The most influential variables in these models could potentially serve as biomarkers or therapeutic targets for the disease.

The goal of this internship is to use counterfactual interpretation [2][3][4][5] of a deep neural network in order to identify relevant biomarkers

The first step will be to construct a highly accurate predictive model for a specified phenotype (such as cancer type or prognosis) utilizing genomic data. Subsequently, we will explore counterfactual explanations for the model's predictions. This process will involve solving constrained optimization problem, integrating the distinct characteristics of genomic data. By comparing real and counterfactual patients, we intend to identify potential biomarkers.

Bibliography

[1] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. ACM Comput. Surv., 51(5):93:1–93:42, Aug. 2018. ISSN 0360-0300.

[2] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR,” SSRN Journal, 2017.

[3] S. Dandl, C. Molnar, M. Binder, and B. Bischl, “Multi-Objective Counterfactual Explanations,” 2020.

[4] S. Verma, J. Dickerson, and K. Hines, “Counterfactual Explanations for Machine Learning: A Review.” arXiv, 2020.

[5] A. Van Looveren and J. Klaise, “Interpretable Counterfactual Explanations Guided by Prototypes.” arXiv, 2020.