Data Science and Artificial Intelligence for Life Sciences at the University of Trieste

Giulia Barbati 1, Luca Bortolussi 1, Giulio Caravagna 2

1) Artificial Intelligence and Cyber Physical Systems Laboratory
2) Cancer Data Science Laboratory

Department of Mathematics and Geosciences, University of Trieste
gbarbati@units.it lbortolussi@units.it gcaravagna@units.it

Abstract

This document reports Data Science and Artificial Intelligence initiatives oriented towards Life Sciences applications that are in place at the Trieste node for the national AIIS Laboratory.

1 Membership

The Trieste node for the AIIS Laboratory involves more than 10 members that apply Data Science and Artificial Intelligence to Life Science. The current members rank as Full Professors, Associate Professors, Tenure-track researchers, Post-Docs and PhD students, and are affiliated to the following departments:

- Mathematics and Geosciences (www.dmg.units.it);
- Engineering and Architecture (https://dia.units.it/it);
- Medical Sciences (https://dsm.units.it/).

In this document we describe relevant ongoing research projects, as well as teaching in the areas of AI,/Data Science for Life Sciences and Healthcare applications.

2 Research groups

2.1 Cancer Data Science Laboratory

The Cancer Data Science (CDS) Laboratory is led by Giulio Caravagna, and consists of computational scientists with different expertises and backgrounds, spanning from computer science, physics, genetics and biology.

The lab develops novel AI technologies for both bulk and single-cell cancer computational data analysis, which are used to study tumours evolution and response to therapy (www.caravagnalab.org).

The CDS Lab is primarily interested in theoretical and applied Data Science for the broad areas of Computational Oncology and Bioinformatics. Among our contributions (https://github.com/caravagnalab) there are model-based technologies i) to measure clonal evolution from high-throughput sequencing data [1,2], ii) to detect repeated evolutionary trajectories with prognostic power [3], and iii) to determine the temporal ordering of somatic mutations from cross-sectional cancer genomics assays [4–6].

Recently, the laboratory has been working on the definition of AI-based models that support the application of whole-genome sequencing technologies at scale [7], as well single-cell RNA sequencing [8]. Other projects are ongoing regarding the application of AI to integrate multi-omics data and, more in general, spatio-temporal measurements.

The CDS Laboratory has direct collaborations with several research institutes that work on AI-related projects, including Human Technopole (IT), the Institute of Cancer Research (UK) the Barts Cancer Institute (UK), University College London (UK), Milan-Bicocca (IT) and Memorial Sloan Kettering (US). The Laboratory also collaborates with several clinical institutes in Italy and abroad: the Royal Marsden Hospital (UK), the Hospital San Raffaele (IT), CRO Aviano (IT) and San Gerardo Monza (IT).

2.2 Artificial Intelligence and Cyber Physical Systems Laboratory and Biostatistics Unit

The Artificial Intelligence and Cyber Physical Systems Lab (AI-CPS) is led by Prof. Luca Bortolussi, and develops novel approaches based on artificial intelligence and machine learning with applications in medicine, in collaboration with the Biostatistics Unit of the Department of Medical Science of the University of Trieste, led by prof. Giulia Barbati. The lab is particularly interested in explainable AI-based techniques for the analysis of biomedical signals of pulmonary ventilation in intensive care and ECG.

The activity of the lab in the area of biomedical signals of assisted ventilatory respiration started several years ago in collaboration with the Intensive Care Unit of the Trieste

University Hospital, and has led to an international patent for the detection of asynchronies between machine and patient efforts.

The lab is also collaborating with the Biostatistics Unit of the Department of Medical Sciences for the application of deep learning to prediction tasks involving ECG signals. Within the same collaboration, the lab has also a growing interest in solutions for high quality synthetic medical data generation based on deep learning generative models. Moreover, there are a series of ongoing projects related to extracting useful epidemiological knowledge by means of machine learning techniques exploiting the informational content of the regional health administrative data.

3 Relevant research project

We describe relevant research projects and collaboration with other academic or industrial partners.

3.1 AI for clonal evolution under therapy

Through an AIRC funded My First AIRC grant (PI Giulio Caravagna) the CDS Laboratory uses AI-based statistical models to study clonal evolution and response to therapy in different types of leukemia.

This collaborative project involves the participation of the Centre of Omics Sciences of Hospital San Raffaele (Dr Giovanni Tonon), which provides sequencing expertise and wet-lab support, and two clinical units based at IRCCS hospitals to provide leukemia samples: the unit of Experimental Onco-hematology at CRO Aviano (Dr Valter Gattei), and the unit of Immunogenetics, Leukemia Genomics and Immunobiology of Hospital San Raffaele (Dr Luca Vago).

The project seeks to develop AI technologies that can better elucidate disease dynamics and relapse mechanisms in both Acute Myeloid Leukemia and Chronic Lymphocytic Leukemia.

3.2 AI for large scale whole-genome sequencing

The CDS Laboratory collaborates synergistically with Genomics England (https://www.genomicsengland.co.uk/), the NHS-owned company that delivers whole-genome sequencing (WGS) to the clinic in the UK.

Genomics England implements, through a collaboration with Illumina (https://www.illumina.com/), large-scale UK-wide genomics projects involving both patients with genetic diseases, and cancer. Giulio Caravagna from the CDS Laboratory is involved in specialised data analysis groups that study WGS data collected for colorectal, endometrial, glioblastoma and haematological cancers.

3.3 Prediction of patient-machine asynchronies in assisted ventilation

The AI-CPS lab has been active for several years in the analysis of respiratory signals (pressure, flow, volume) returned by pulmonary ventilators, particularly when they are used in assisted respiration mode [9]. In this case, the ventilator responds to a trigger of the patient asking for air by a positive pressure that helps the patient to breathe. Sometimes, the trigger mechanism can be misinterpreted by the controller of the machine, resulting in a so-called asynchrony. While a single asynchrony is of limited concern, several repeated episodes can lead to long term damages of the respiratory ability. However, detection of asynchronies is an error prone, manual activity. We developed a method, patented, that combines geometrical methods to extract features from the signal pipelined with machine learning which has a very high specificity and sensitivity [10]. We are currently working on extensions of this method and application of AI approaches for the detection of other issues related to patients under assisted and forced ventilation, also using multi-modal learning strategies, combining different kinds of signals.

3.4 Deep Learning methods in Medicine

In collaboration with the Biostatistics unit of the Department of Medical Sciences and with the cardiology medical unit, we are exploring the use of deep learning approaches for the analysis of ECG signals in order to predict the onset of cardiovascular diseases. In particular, we investigated the performances of two ML approaches based on ECGs for the prediction of new-onset atrial fibrillation (AF), in terms of discrimination, calibration and sample size dependence [11]. Currently, we are extending the analysis to the time-to-event framework to explore the relationship between predictive accuracy and time distance from ECG recording.
Within the survival analysis framework, we implemented a deep-learning-based prognostic model for incident heart failure (HF) in patients with diabetes using electronic health records (EHR) that takes into account a large and heterogeneous set of clinical factors [12]. Our results suggest that prognostic models may improve using EHRs in combination with AI techniques for survival analysis, which provide high flexibility and better performance with respect to standard approaches.

3.4 AI for Infectious Disease Control

Dr Alberto d'Onofrio, leader of the newly formed Computer Science for Complex Systems Laboratory, works at the interface between AI and the modelling of biological systems, with particular focus on Computational Epidemiology of Infectious Diseases and related problems in

Global Public Health. He is currently co-leading the team writing a report for WHO on this and related matters.

3.5 AI for Radiobiochemistry

Dr Alejandro Rodriguez-Garcia is currently working on the application of AI to the understanding of the biological availability of radioactive markers. In concrete, Unsupervised Machine Learning techniques are applied to analyze molecular simulations of some markers and obtain a deeper understanding that could lead to the proposal of new active substances with better biological properties.

3.5 AI for Neurosciences

Dr Fabio Anselmi is currently working at the intersection between computational neuroscience and machine learning.
In particular he develops machine learning models of the visual cortex that embody biological constraints and priors on how humans understand visual scenes such as the concept of whole-and parts in object recognition
He also develops models that incorporate symmetries of the visual signal statistics into the processing pipeline such as transformations that preserve object classification with the aim to have more faithful models of how the visual cortex works

4 Teaching activities

We offer initiatives at the crossings of AI and Machine Learning applied to Life Sciences at several undergraduate and postgraduate levels. Most specialized teaching is concentrated in the Data Science and Scientific Computing (https://dssc.units.it/) Master program and in the Applied Data Science and AI doctoral program (http://adsai.units.it).

Relevant courses to the health area are:

- Genome Data Analytics (6 CFU), taught by Prof. Giulio Caravagna, covering topics on the application of Machine Learning to genome-sequencing assays, with a strong focus on cancer data analysis.

- Health Data Analytics (6 CFU), taught by Prof. Giulia Barbati, covering topics at the crossings between biostatistics and Machine Learning, with application to medical tabular data and survival analysis.

Riferimenti bibliografici

1. Caravagna G, Sanguinetti G, Graham TA, Sottoriva A. The MOBSTER R package for tumour subclonal deconvolution from bulk DNA whole-genome sequencing data. *BMC Bioinformatics*. 2020;21(1):531.

2. Caravagna G, Heide T, Williams MJ, Zapata L, Nichol D, Chkhaidze K, Cross W, Cresswell GD, Werner B, Acar A, Chesler L, Barnes CP, Sanguinetti G, Graham TA, Sottoriva A. Subclonal reconstruction of tumors by using machine learning and population genetics. *Nat Genet*. 2020;52(9):898-907.

3. Caravagna G, Giarratano Y, Ramazzotti D, Tomlinson I, Graham TA, Sanguinetti G, Sottoriva A. Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nat Methods*. 2018;15(9):707-714.

4. Olde Loohuis L, Caravagna G, Graudenzi A, Ramazzotti D, Mauri G, Antoniotti M, Mishra B. Inferring tree causal models of cancer progression with probability raising. *PLoS One*. 2014;9(10):e108358.

5. Caravagna G, Graudenzi A, Ramazzotti D, Sanz-Pamplona R, De Sano L, Mauri G, Moreno V, Antoniotti M, Mishra B. Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proc Natl Acad Sci U S A*. 2016;113(28):E4025-E4034.

6. De Sano L, Caravagna G, Ramazzotti D, Graudenzi A, Mauri G, Mishra B, Antoniotti M. TRONCO: an R package for the inference of cancer progression models from heterogeneous genomic data. *Bioinformatics*. 2016;32(12):1911-1913.

7. Househam J, Bergamin R, Milite S, Cross WCH, Caravagna G. Integrated quality control of allele-specific copy numbers, mutations and tumour purity from cancer whole genome sequencing assays. *biorXiv*. Published online 2021. doi:10.1101/2021.02.13.429885

8. Milite, Salvatore, Riccardo Bergamin, Lucrezia Patruno, Nicola Calonaci, and Giulio Caravagna. "A Bayesian method to cluster single-cell RNA sequencing data using copy number alterations." Bioinformatics 38, no. 9 (2022): 2512-2518.

9. Casagrande A, Quintavalle F, Fernandez R, Blanch L, Ferluga M, Lena E, Fabris F, Lucangelo U. An effective pressure–flow characterization of respiratory asynchronies in mechanical ventilation. *J Clin Monit Comput*. 2021;35(2):289-296.

10. Bufo S, Bartocci E, Sanguinetti G, Borelli M, Lucangelo U, Bortolussi L. Temporal Logic Based Monitoring of Assisted Ventilation in Intensive Care Patients. In: *Leveraging Applications of Formal Methods, Verification and Validation. Specialized Techniques and Applications*. Springer Berlin Heidelberg; 2014:391-403.

11. Giovanni Baj, Ilaria Gandin, Arjuna Scagnetto, Luca Bortolussi, Chiara Cappelletto, Andrea Di Lenarda, Giulia Barbati, Machine learning approaches for ECG-based models: discrimination and calibration for atrial fibrillation prediction, 16 February 2023, Submitted.

Preprint available at Research Square [https://doi.org/10.21203/rs.3.rs-2509748/v1]

12. Gandin I, Saccani S, Coser A, Scagnetto A, Cappelletto C, Candido R, Barbati G, Di Lenarda A. Deep-learning-based prognostic modeling for incident heart failure in patients with diabetes using electronic health records: A retrospective cohort study PLOS ONE 18(2): e0281878 (2023)