3iA Côte d'Azur
Interdisciplinary Institute
for Artificial Intelligence

Ph.D. research topic

- Title of the proposed topic:

  Data Structuration and Security in Large-Scale Collaborative Healthcare Data Analysis

- Research axis of the 3IA: Axis 2
- Supervisor (name, affiliation, email):

  - Melek Önen - EURECOM, Sophia Antipolis - melek.onen@eurecom.fr
  - Olivier Humbert - Antoine Lacassagne Center; TIRO laboratory (UMR E 4320, CEA/UCA) - Olivier.HUMBERT@univ-cotedazur.fr
  - Marco Lorenzi - EPIONE, Inria Sophia Antipolis - marco.lorenzi@inria.fr

- Potential co-supervisor (name, affiliation):
- The laboratory and/or research group:
  - EURECOM, Sophia Antipolis
  - Antoine Lacassagne Center; TIRO laboratory
  - EPIONE, Inria Sophia Antipolis

Apply by sending an email directly to the supervisor.
The application will include:

- Letter of recommendation of the supervisor indicated above
- Curriculum vitæ.
- Motivation Letter.
- Academic transcripts of a master's degree(s) or equivalent.
- At least, one letter of recommendation.
- Internship report, if possible.

- Description of the topic:

Real-world applications of artificial intelligence must often adhere to strict constraints concerning the non-transferability of sensitive information across centres. These limitations impose important methodological challenges for the application of current powerful learning methods to healthcare data. Furthermore, when machine learning is based on agglomerated multi-centric data, the problem of bias and fairness becomes critical. In particular, model's prediction and parameters may importantly differ when trained on multiple data instances characterized by different conditions, which can either be known, such as missing data, data acquisition bias and heterogeneity in clinical cohorts across hospitals, or unknown. For all these reasons, the problem of security and fairness in machine learning is nowadays a central issue when deploying modern and large-scale learning systems to healthcare scenarios.

Federated learning (FL) [1] offers the opportunity of securely training models on private information distributed across data centres, without the need to centrally store the data. In particular, our current project Fed-BioMed is aiming at establishing a federated learning infrastructure for the deployment of IA in multi-centric hospital data, while guaranteeing the security for the data stored at each clinical centre [2,3,4]. Our current application targets the prediction of response to immunotherapy from the analysis of 3D PET/CT images and clinical data [5-6]. Our project will leverage on the clinical datasets provided by a network of participating hospitals from the UNICANCER consortium (Comprehensive Cancer Centres from Nice, Paris, Lyon, Caen, Rouen, Toulouse, Rennes). The dataset pre-processing and structuration in a homogeneous way across hospitals is a crucial step for the Federated Learning approach.

This research scenario provides a unique setup for the investigation of novel approaches and instruments for implementing clean and structure medical datasets and for guaranteeing security and fairness in large-scale collaborative healthcare data analysis. In particular, this project will focus on the methodological developments necessary to fill the existing gap for the effective exploitation of FL and IA in multi-centric hospital data. We will focus on the following issues:

1) Learning harmonisation strategies for medical imaging and clinical data preparation in Federated Learning. Multi-centric hospital data requires extensive manipulation and verification in order to be adequately prepared and structured before being analysed in machine learning process. This kind of manipulation spans from data imputation to bias removal and detection of coherent patterns of variability. In most of applications, this problem is often underestimated and tackled in ad ad-hoc fashion, mostly depending on manual operations.

This part of the project will be carried out in close collaboration with the medical team and the Information System Management direction (DSI) of the Antoine Lacassagne Cancer Center (Nice, France). The DSI are currently deploying automatic process for medical imaging data extraction and structuration. The PhD student will collaborate to the elaboration of a warehouse of heterogeneous "cleaned" and structured data in the hospital, in order to guarantee accurate data-preparation for the federated analysis.

We will focus on a coherent modelling framework targeting the following problems:

- *Account for missing data patterns, either in a missing at random (MAR) and not at random (MNAR) setting.* These tasks require the definition of data imputation techniques which could scale to high-dimensional and heterogeneous data. At the same time, we will focus on the development of novel federated optimization strategies allowing to account for heterogeneity of features and data modalities across centres.
- *Accounting for different sources of data bias which affecting high-dimensional data.* In this setting, classical bias and standardization techniques (such as based linear mixed effect modelling) do not scale properly and are highly suboptimal to account for complex data variation. We envisage the extension of current dimensionality reduction approaches based on latent variable modelling, such as probabilistic PCA or variational autoencoding, to account for different components of variability. This approach requires the specification of a random effect structure in current latent variable models, through the learning of the latent components associated to respectively signal, bias confounders and noise.

2) Studying the impact of privacy-preserving methods in FL for biomedical applications. Due to the highly sensitive nature of the medical data being processed, we also plan to investigate the use or design of privacy preserving primitives for dedicated operations performed with federated learning.

Primitives ensuring differential privacy (DP) [7] consist of adding some noise up to a certain bias to the raw data. The choice of the differentially private protection mechanism should not have a strong impact on the actual utility of the underlying learning algorithm. Hence, one should study the trade-off between the privacy level and the accuracy of the obtained model. On the other hand, while differentially private federated learning may achieve a certain level of privacy, the underlying data still remain in the original form, and processing such critical data in raw format (even though differentially private) could still be avoided. We therefore intend to make use of advanced cryptographic tools such as homomorphic encryption (HE) [8] or secure multi-party computation (MPC) [9] which allow the processing of the data without the leakage of additional information. While there exist several studies on the use of such cryptographic tools in machine learning (see [10] for example), little is known about the impact of these privacy preserving primitives when applied to complex data, such as medical images and high dimensional biological measurements and their application to a multi-party setting. Moreover, the coherent integration of DP and cryptographic tools such as MPC or HE in federated learning schemes requires the identification of optimal data and novel strategies for not significantly compromising efficiency and accuracy of the learning process,

Bibliographical references:

[1] H. Brendan McMahan, Eider Moore, Daniel Ramage, et al. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), 2017.

[2] Santiago Silva, Boris Gutman, Barbara Bardoni, Paul M Thompson, Andre Altmann, Marco Lorenzi. Federated Learning in Distributed Biomedical Databases: Meta-analysis of Large-scale Brain Imaging Data. IEEE International Symposium on Biomedical Imaging (ISBI), Venice, 2019.

[3] Santiago Silva, Andre Altmann, Boris Gutman, Marco Lorenzi. Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data. In Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning (pp. 201-210). Springer, Cham. 2020.

[4] Yann Fraboni, Richard Vidal and Marco Lorenzi. Free-rider Attacks on Model Aggregation in Federated Learning. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), 2021.

[5] Humbert O et al. (18)FDG PET/CT in the early assessment of non-small cell lung cancer response to immunotherapy: frequency and clinical significance of atypical evolutive patterns. J. Eur J Nucl Med Mol Imaging. 2020 May;47(5):1158-1167.

[6] Chardin D, Paquet M, Schiappa R, Darcourt J, Bailleux C, Poudenx M, Sciazza A, Ilie M, Benzaquen J, Martin N, Otto J, Humbert O. Baseline metabolic tumor volume as a strong predictive and prognostic biomarker in patients with non-small cell lung cancer treated with PD1 inhibitors: a prospective study. J Immunother Cancer. 2020 Jul;8(2):e000645.

[7] C. Dwork, Differential privacy, International Colloquium on Automata, Languages and Programming (ICALP), 2006.

[8] C. Gentry, Computing Arbitrary Functions on Encrypted Data, Communications of the ACM, Vol. 53, No. 3, pages 97-105, March 2010.

[9] D. Evans, V. Kolesnikov, M. Rosulek, A Pragmatic Introduction to Secure Multi-Party Computation, NOW Publishers, 2018.

[10] M. Azraoui, M. Bahram, B. Bozdemir, S. Canard, E. Ciceri, O. Ermis, R. Masalha, M. Mosconi, M. Önen, M. Paindavoine, B. Rozenberg, B. Vialla, S. Vicini. SoK: Cryptography for Neural Networks, IFIP Summer school on Privacy and Identity Management, 2020.