Evaluating Novel Quality Measures for Explainers

The position is a 3 years fully funded PhD, starting October 1st 2025.

Keywords: Inherently Interpretable Models, Causality, Multivariate Time Series, Explaination Quality

1 Context

The AIDA project aims to address fundamental issues related to the quality and the applicability of explanations produced for Deep Learning (DL) models driven by the recent vision of actionable explainable AI (aXAI) [2]. AIDA considers so-called actionable concepts, measures, and metrics for explainable learning and reasoning. In particular, it focuses on more expressive forms of explanations that can answer not only why questions (why do we obtain a specific prediction, given the features of input observations?) but also action-guiding explanations such as how-to (what are the necessary actions to change the prediction of a specific input observation?) and what-if (what are the necessary and minimal sets of actions on input observations required to obtain an alternative prediction?). Answers to these questions are crucial to act on the models and the data used in various prediction tasks of real applications [1].

2 Goals

The PhD thesis will focus on interactive exploration of explanations. To this aim, several research questions will be addressed. One first objective of this PhD is to develop and implement new quality measures for evaluating the effectiveness and actionability of explanations in multivariate time series (MTS) predictive models. This research will focus on creating metrics that assess the plausibility, diversity, and actionability of explanations, as well as alignment with domain knowledge and user expectations. The second objective will be to propose new ways of exploring explanation space based on the previous metrics. Finally, we will ensure the faithfulness of explanations by proposing novel approaches to align explanation exploration with expert knowledge.

3 Advising and application

Employers: University of Tours (France)

Labs: Laboratoire d'Informatique Fondamentale et Appliquée de Tours (LIFAT) and Institut de Recherche en Informatique de Toulouse (IRIT)

Locations: Tours (France) : The candidate will be located in Tours. Several research visits to Toulouse site (expenses fully covered) are expected during the PhD.

Supervisors:

- Nicolas Labroche, Professor, University of Tours, labroche@univ-tours.fr,
- Julien Aligon, Associate Professor HDR, Toulouse Capitole University, julien.aligon@irit.fr,
- Moncef Garouani, Associate Professor, Toulouse Capitole University, moncef.garouani@irit.fr,
- Alexandre Chanson, Associate Professor, University of Tours, chanson@univ-tours.fr,

Requirements: Applicants are expected to hold a Master's degree in Computer Science, be skilled in machine learning, programming and be fluent in English. Applicants must demonstrate proficiency in one of the following topics: **machine learning** with an emphasis on inherently interpretable models, **data analysis**, **statistics**. Applications failing to match these expectations to will be automatically rejected.

Application: This position is opened **until filled** or **before June 2, 2025** (firm deadline).
Applicants will email, the following documents to the supervisors: CV, transcripts of the Master's program, Master's thesis dissertation, cover letter, reference letters.

Shortlisted applicants will be contacted for a video interview that will include a discussion of the scientific literature relevant to the topic.

References

[1] Sander Beckers. Causal explanations and XAI. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 90–109. PMLR, 11–13 Apr 2022.

[2] Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek. *xxAI - Beyond Explainable Artificial Intelligence*, pages 3–10. 04 2022.