Université de technologie de Compiègne – Thesis proposal

Part 1: Scientific sheet	
Thesis proposal title	**Robustness in Machine Learning Explanations**
Financial resources	French National Research Agency - ANR
Host laboratory	Research team: CID, Heudiasyc UMR 7253
Thesis supervisors	Vu-Linh NGUYEN (Junior professor chair) Sébastien DESTERCKE (CNRS senior researcher) Mylène MASSON (Associate professor, HDR)
Scientific domain(s)	Computer science
Research work	**In a sentence, this thesis aims to extend explainable AI methods to make them more robust and increase the trustworthiness of the system. The supervising team has strong expertise in robust AI and machine learning in general, with a specific focus on uncertainty/robustness quantification methods.** Despite an increasingly large body of literature in the field of explainable AI [3, 6, 8, 9], the evaluation of explainable methods remains a challenging problem [4]. The difficulty of the evaluation task is typically introduced by a combination of different factors, including but not limited to the lack of ground-truth explanations, the unreliability of the predictions, which is often a consequence of model inadequacy and/or data imperfections (in terms of quality and/or quantity) and may lead to uninformative explanations, the non-uniqueness of predictions (due to random factors of the model [5, 7]), which can lead to the non-uniqueness of explanations [1]. Moreover, even if the prediction is unique, explanation methods may produce unstable explanations, i.e., negligibly small perturbations to an instance can result in substantially different explanations, and non-unique explanations, i.e., multiple runs on the same input instance with the same parameter settings may result in vastly different explanations [10]. This project is devoted to the development of modeling and quantifying the robustness of explanation methods and their applications in constructing robust explanation methods. The first aim of the project, i.e., modeling and quantifying the robustness of explanation methods, would directly facilitate the evaluation task. The second one, i.e., constructing robust explanation methods, would beneficially enlarge the existing set of explanation methods. Methodologically, we treat the random and perturbated factors as sources of uncertainty/unrobustness and develop methods to quantitatively model the robustness of explanations under the presence of these factors. The candidate is encouraged to start with commonly used explanation methods, such as **SHAP** [**8**], **LIME** [**9**] and counterfactual explanations [3], and intuitive and commonly used predictive models, such as **tree-based models** [**2, 6**], to gradually gain the relevant expertise, when basing her/his results, software and experimental protocols, and to communicate them through scientific articles. Depending on the progress, we can then look at other commonly used predictive models, such as **Bayesian neural networks** [**5**] and **Monte Carlo dropout predictions** [**7**]. **To achieve this, the candidate will join a growing team supported by two chairs (SAFE AI and Trustworthy AI junior professor chair), benefiting from the associated environment.**
Starting time	As soon as possible
Duration	36 months
Keywords	Uncertainty quantification, Accelerated machine learning, Trustworthy AI

Part 2: Job description	
Requirements	Master 2 or engineer in computer science with good skills in statistics and data mining, and/or good programming skills (Python, PyTorch, TensorFlow, ...). Experience with explainble AI toolkits (Quantus, InterpretDL, OmniXAI, ...) is a plus.
Additional missions	Teaching is possible, but not mandatory
Research laboratory	Heudiasyc UMR 7253, Université de Technologie de Compiègne
Material resources	Shared office, laptop, access to the laboratory’s GPU servers and the Jean Zay supercomputer installed at IDRIS, as well as to the laboratory’s platforms, ...
Human resources	Internal and external collaborations
Working conditions	The candidate is funded by French National Research Agency - ANR and shall be provided with financial support for traveling (conferences, workshops, summer schools, short-term visits, ...)
Research project	Trustworthy AI Chair, SAFE AI Chair
National collaborations	
International collaborations	UAI team, Eindhoven University of Technology, The Netherlands.
International co-supervision	No
Contact	Applications and questions can be sent to: - Vu-Linh Nguyen (vu-linh.nguyen@hds.utc.fr) - Sébastien Destercke (sebastien.destercke@hds.utc.fr) - Mylène Masson (mylene.masson@hds.utc.fr)

Applicant files

Applications must include the following items:

- a letter of motivation detailing explicitly what is the interest of the applicant in the proposed topic;
- a curriculum vitae which clearly shows how the candidate profile matches the above requirements and highlights how the candidate’s experience relates to the proposed topic;
- contact information of at least one reference (two or more would be appreciated).
- transcripts and existing theses;

Any application not containing these items, or not tailored to this proposal, will not be considered further. In addition, the following optional items may be included:

- existing scientific papers;
- any link to significant realisations (e.g., software, . . .).

References

[1] K. Bykov, M. M.-C. Höhne, A. Creosteanu, K.-R. Müller, F. Klauschen, S. Nakajima, and M. Kloft. Explaining bayesian neural networks. *arXiv preprint arXiv:2108.10346*, 2021.

[2] L. Grinsztajn, E. Oyallon, and G. Varoquaux. **Why do tree-based models still outperform deep learning on typical tabular data?** In *Proceedings of the Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

[3] R. Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022.

[4] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, and M. M.-C. Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.

[5] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun. **Hands-on Bayesian neural networks—A tutorial for deep learning users**. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022.

[6] A. Karczmarz, T. Michalak, A. Mukherjee, P. Sankowski, and P. Wygocki. **Improved Feature Importance Computation for Tree Models Based on the Banzhaf Value**. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2022.

[7] A. Lemay, K. Hoebel, C. P. Bridge, B. Befano, S. De Sanjosé, D. Egemen, A. C. Rodriguez, M. Schiffman, J. P. Campbell, and J. Kalpathy-Cramer. **Improving the repeatability of deep learning models with Monte Carlo dropout**. *npj Digital Medicine*, 5(1):174, 2022.

[8] S. M. Lundberg and S.-I. Lee. **A unified approach to interpreting model predictions**. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, pages 4768–4777, 2017.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin. **"Why should i trust you?" Explaining the predictions of any classifier**. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD)*, pages 1135–1144, 2016.

[10] D. Z. Slack, S. Hilgard, S. Singh, and H. Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, pages 9391–9404, 2021.