Université de technologie de Compiègne – Thesis proposal

Part 1: Scientific sheet	
Thesis proposal title	**Explanation Methods for Uncertainty Estimates**
Financial resources	French National Research Agency - ANR
Host laboratory	Research team: CID, Heudiasyc UMR 7253
Thesis supervisors	Vu-Linh NGUYEN (Junior/Assistant professor) 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 in the system. The supervising team has very strong expertise in robust AI and machine learning in general, with a specific focus on uncertainty quantification methods.** The main goal of many explanation methods, such as **SHAP** [7], LIME [15] and counterfactual explanations [2], is to produce (human-)interpretable descriptions of model predictions. Such interpretable descriptions mainly serve the purpose of explaining "why the model makes its predictions", which in turn mainly serves the purpose of model auditing (i.e., to validate/debug the model predictions). Under the presence of uncertainty, which is often a consequence of model inadequacy and/or data imperfections (in terms of quality and/or quantity), the model can however be uncertain about its predictions and makes unreliable predictions. Thus, the robustness and usefulness of associated explanations can be questioned consequently. Explanation methods which seek **interpretable descriptions of uncertainty associated with model predictions** have received increasing attention in recent years [1, 13]. Such explanation methods can serve multiple purposes, such as explaining "why the model is uncertain about its predictions", explaining "when and why making cautious/preference predictions may be more beneficial than making precise predictions" and suggesting treatments (such as making reliable cautious/preference predictions, active feature selection for enriching available data, and providing guidance for revising the hypothesis space) which can improve the predictive system. It is clear that the development of explanation methods for (predictive/model) uncertainty estimates can not be decoupled from the development of methods for uncertainty modelling and quantification. Therefore, there is an obvious need for developing explanation methods which take into account recent advances in uncertainty modelling and quantification (e.g., [3, 4]) and its corresponding treatments (making cautious predictions [9], learning preference orders [11], acquiring additional data [12], etc.). This project aims to develop explanation methods which seek interpretable descriptions of uncertainty associated with model predictions. In particular, it shall focus on explaining different sources of uncertainty, such as **epistemic uncertainty and aleatoric uncertainty** [3]. Another topic of interest is uncertainty quantification in the explanation themselves, in order to make robust explanations and to explore the use of interpretable descriptions as augmented information for **making (cautious/preference) predictions** [9, 11]. Once methods are developed, we want to assess its usefulness in high stakes applications, such as **healthcare data analysis and autonomous vehicles**. **To achieve this, the student will join a growing team supported by two chairs (SAFE AI and Trustworthy AI junior professor chair), benefiting from the associated environment.**

Starting time	01/10/2023
Duration	36 months
Key words	XAI, Uncertainty quantification, Robust explainability, Trustworthy AI

Part 2: Concrete Problems	
In the following, we summarize 2 concrete problems which would serve as the point of departure.	
Uncertainty-informed classifiers	In classification, several feature space partitioning classifiers (FSPCs), which seek a bias-variance tradeoff, have been proposed. Instead of learning a global classifier, it partitions the feature space into a number of (possibly overlapping) regions and learns local region-specific classifiers. For example, decision trees [5] partition the feature space into box shaped regions/leaves and learn a set of voting classifiers, one per region. A stronger discriminative ability has been sought by equipping regions with parametric classifiers [6]. The assumption of having box shaped regions can be relaxed by using appropriate partitioning/splitting rules, such as reject classifiers [18] and clustering techniques [17]. Image classification data sets can be tackled by equipping each region with a convolutional neural network (CNN) [10]. FSPCs would be in principle generalized to handle mixed data by partitioning/splitting the (continuous) feature space using discrete features and restricting the domain of local classifiers to continuous feature space [19]. FSPCs can also inform uncertainty estimates by equipping each region with an uncertainty-informed classifier [8, 20]. To make contribution in this line of research, the first part of the thesis will focus on developing FSPCs whose local region-specific classifiers inform quantitative information of different sources of uncertainty. Examples of such uncertainty-informed classifiers are **reliable classifiers [11, 16], evidential classifiers [14] and Bayesian neural networks (BNNs) [4]**.
Explanation Methods	The second part of the thesis will be devoted to the development of **explanation methods for (either the original or FSPC version of) uncertainty-informed classifiers** (such as reliable classifiers [11, 16], evidential classifiers [14] and BNNs [4]).

Part 3: 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 toolkits for interpretation algorithms (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 supports for travelling (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 are 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 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] J. Antoran, U. Bhatt, T. Adel, A. Weller, and J. M. Hernández-Lobato. **Getting a CLUE: A Method for Explaining Uncertainty Estimates**. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*, 2021.

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

[3] E. Hüllermeier and W. Waegeman. **Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods**. *Machine Learning*, 110:457–506, 2021.

[4] A. Kendall and Y. Gal. **What uncertainties do we need in Bayesian deep learning for computer vision?** In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, pages 5580–5590, 2017.

[5] S. B. Kotsiantis. Decision trees: a recent overview. *Artificial Intelligence Review*, 39:261–283, 2013.

[6] N. Landwehr, M. Hall, and E. Frank. Logistic model trees. *Machine learning*, 59:161–205, 2005.

[7] 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.

[8] C. J. Mantas and J. Abellan. Credal-c4. 5: Decision tree based on imprecise probabilities to classify noisy data. *Expert Systems with Applications*, 41(10):4625–4637, 2014.

[9] T. Mortier, M. Wydmuch, K. Dembczyński, E. Hüllermeier, and W. Waegeman. **Efficient set-valued prediction in multi-class classification**. *Data Mining and Knowledge Discovery*, 35(4):1435–1469, 2021.

[10] V. N. Murthy, V. Singh, T. Chen, R. Manmatha, and D. Comaniciu. Deep decision network for multi-class image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2240–2248, 2016.

[11] V.-L. Nguyen, S. Destercke, M.-H. Masson, and E. Hüllermeier. **Reliable multi-class classification based on pairwise epistemic and aleatoric uncertainty**. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5089–5095, 2018.

[12] V.-L. Nguyen, M. H. Shaker, and E. Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1):89–122, 2022.

[13] I. Perez, P. Skalski, A. Barns-Graham, J. Wong, and D. Sutton. **Attribution of Predictive Uncertainties in Classification Models**. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2022.

[14] B. Quost, M.-H. Masson, and S. Destercke. **Dealing with atypical instances in evidential decision-making**. In *Proceedings of the 14th International Conference on Scalable Uncertainty Management (SUM)*, volume 12322, pages 217–225, 2020.

[15] 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.

[16] R. Senge, S. Bösner, K. Dembczyński, J. Haasenritter, O. Hirsch, N. Donner-Banzhoff, and E. Hüllermeier. **Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty**. *Information Sciences*, 255:16–29, 2014.

[17] T. Stepišnik and D. Kocev. Oblique predictive clustering trees. *Knowledge-Based Systems*, 227:107228, 2021.

[18] J. Wang and V. Saligrama. Local supervised learning through space partitioning. In *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPs)*, pages 91–99, 2012.

[19] Y. Yang. Generalized bayesian network classifiers. Master's thesis, Eindhoven University of Technology, the Netherlands, 2022.

[20] H. Zhang, B. Quost, and M.-H. Masson. Cautious random forests: a new decision strategy and some experiments. In *Proceedings of the 12th International Symposium on Imprecise Probability: Theories and Applications (ISIPTA)*, 2021.