Exploring the interplay between uncertainty quantification and interpretability of machine learning models

Dr. Nikos Nikolaou (P&A), Prof. Ingo Waldmann (P&A)

This project will explore the following key question: *'Can we make machine learning models more interpretable by improving how they quantify uncertainty over their predictions?'*

Machine learning (ML) algorithms are driving innovation across domains, Exoplanetary Science [1-4] and Biomedicine [5-8] among them. Two important yet often overlooked aspects of machine learning models are *interpretability* [9, 10] and *uncertainty quantification* [11, 12].

Most popular ML algorithms (e.g. deep neural networks, ensembles), are notorious for being 'black boxes', but several *model interpretability methods* -of various degrees of reliability [9, 10]- have been developed to allow us to understand their inner workings, to uncover hidden biases in the models (or the data), to increase trustworthiness and adoption, to inspect when and how they fail, or to uncover new domain knowledge. For instance, *feature importance methods*, attempt to identify the key features that drive model predictions. Some of the open challenges related to these techniques include handling the non-uniqueness of explanations of model behaviour ('*Rashomon effect*'), reliably quantifying uncertainty over feature attribution and quantitatively evaluating them.

Similarly, popular ML algorithms are also known to suffer from poor uncertainty estimation over their predictions [11, 12]. Correctly quantifying uncertainty is crucial in the presence of large amounts of data for prioritizing objects for further analysis (common in Exoplanetary Science) and in cost-sensitive applications for balancing costs and benefits for optimal decision making based on model predictions (common in Biomedicine). Regression models often fail to provide cohesive confidence intervals and classification models often ignore class membership probabilities -or obtain systematically unreliable estimates thereof [13, 14]. Techniques to detect and address these issues include *conformal prediction* [16, 17] and *probabilistic calibration* [13, 14, 17, 18].

In this project we will investigate if these two weaknesses are connected and whether by addressing one, we can improve upon the other.

Proposed subtasks for the CDT PhD project (each to take ~1 year, but likely to be explored in parallel):
1. **Explore the effect of improving model uncertainty estimation on predictive performance & interpretability.**
2. **Investigate the use of calibration measures (e.g. Brier Score) of feature importance scores to quantitatively evaluate model interpretability methods, ideally by establishing a single performance measure.**
3. **Use the above performance measure/methodology to (i) design new model interpretability methods and/or to (ii) inform model training to produce more interpretable models.**

We will explore the interplay between uncertainty estimation and model interpretability in the context of ML models for select applications in exoplanetary science[1] and medical imaging[2]. The supervising team has extensive experience in using interpretability methods in exoplanetary science [1, 2, 4] and biomedical [7, 8] applications. This is an ambitious project, the findings of which can inform & advance both ML and each of the application fields.

Both supervisors have extensive experience in the intersection of ML & exoplanetary science (e.g. [1-4]). **Dr. Nikolaou** is a Lecturer at DISI and an expert in ML with an emphasis in model interpretability. He has worked on applications in exoplanetary science, biomedicine & imaging. He is currently supervising 2 CDT students & several MSc projects exploring ML model interpretability & uncertainty quantification in similar settings. His PhD thesis was on uncertainty quantification of gradient boosting classifier ensembles [13]. **Prof. Waldmann** is a Professor at UCL and an expert in exoplanetary science. He has led the ExoAI Team at UCL and has pioneered the use of ML methods in exoplanet detection & characterization. He is currently supervising several PhD students & postdocs in related projects.

[1] An example application is inferring atmospheric characteristics from exoplanetary spectra (exoplanet characterization). A dataset that can be used in this task is the one provided here for the Ariel Data Challenge: https://zenodo.org/record/6770103#.Y2PmuuzP1qs

[2] An example application is classification of histological whole slide images (WSIs). An appropriate dataset will become available by the start of the project.

References:

[1] Yip, K. H., et al. "Pushing the Limits of Exoplanet Discovery via Direct Imaging with Deep Learning". *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Cham, 2019.
[2] Yip, K. H., et al. "Peeking inside the Black Box: Interpreting Deep-learning Models for Exoplanet Atmospheric Retrievals". *The Astronomical Journal* 162.5 (2021): 195.
[3] Morvan, M., et al. (2022). "Don't Pay Attention to the Noise: Learning Self-supervised Representations of Light Curves with a Denoising Time Series Transformer". In *Machine Learning for Astrophysics: Workshop at ICML 2022* (Vol. 162). *arXiv:2207.02777*
[4] Nikolaou, N., et al. (2023). "Lessons Learned from the 1st ARIEL Machine Learning Challenge: Correcting Transiting Exoplanet Light Curves for Stellar Spots". *To Appear in RAS Techniques & Instruments*, *arXiv:2010.15996*
[5] Huff, D. T. et al. (2021). Interpretation and visualization techniques for deep learning models in medical imaging. *Physics in Medicine & Biology*, *66*(4), 04TR01.
[6] Salahuddin, Z., et al. (2022). "Transparency of deep neural networks for medical image analysis: A review of interpretability methods". *Computers in biology and medicine*, *140*, 105111.
[7] Ellen, J. G., et al. (2023). "Autoencoder-based multimodal prediction of non-small cell lung cancer survival. *Scientific Reports*, *13*(1), 15761.
[8] Neocleous Y. & Nikolaou N. (2024) "What model interpretability teaches us about machine learning models for skin lesion classification". *Under review, preprint available upon request.*
[9] Molnar, C. (2020). "*Interpretable machine learning"*. Lulu.com.
[10] Rudin, C., et al. (2022). "Interpretable machine learning: Fundamental principles and 10 grand challenges". *Statistic Surveys*, *16*, 1-85.
[11] Psaros, A. F. et al. (2023). "Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons". *Journal of Computational Physics*, *477*, 111902.
[12] Abdar, M., et al. (2021). "A review of uncertainty quantification in deep learning: Techniques, applications and challenges". *Information fusion*, *76*, 243-297.
[13] Nikolaou, N. (2016). "*Cost-sensitive boosting: A unified approach",* PhD Thesis, The University of Manchester, UK.
[14] Vaicenavicius, J. et al. (2019). "Evaluating model calibration in classification." In *The 22nd International Conference on Artificial Intelligence and Statistics* (pp. 3459-3467). PMLR.
[15] Angelopoulos, A. N., & Bates, S. (2021). "A gentle introduction to conformal prediction and distribution-free uncertainty quantification". *ArXiv preprint, arXiv:2107.07511*.
[16] Vovk, V. et al. (2022). "Algorithmic Learning in a Random World", Springer
[17] Silva Filho, T. et al. (2023). "Classifier calibration: a survey on how to assess and improve predicted class probabilities". *Machine Learning*, 1-50.
[18] Wang, C. (2023). "Calibration in Deep Learning: A Survey of the State-of-the-Art". *ArXiv preprint, arXiv:2308.01222*.