ESSAI-ACAI 2024 Course Proposal: Explainable Machine Learning

Nikos Nikolaou[1]

[1]**Centre for Data Intensive Science & Industry, UCL, Gower Street, WC1E 6BT, London, UK**

ABSTRACT

Machine learning models are often perceived as '*black boxes*'. *Explainable machine learning (XML)* a.k.a. *Explainable artificial intelligence (XAI)* methods allow us to inspect their inner workings and understand their predictions to reveal new insights about the data, but also hidden biases, increase transparency, trust and safety of ML applications and promote their adoption by domain experts and general public alike. In high stakes applications (e.g. medical, safety) XAI is becoming a regulatory requirement. The course will provide an overview of principles, methods, applications, limitations and challenges of XAI. We will cover the more inherently interpretable '*white box*' methods as well as methods tailored to the more opaque deep neural networks and general purpose methods applicable to any model. We will give examples of successful applications, discuss advantages and disadvantages of each approach, but also limitations and open problems of XAI as a field. The course will be accompanied by Python tutorials.

Keywords: Explainable Artificial Intelligence (XAI), Explainable Machine Learning (XML), Interpretable Machine Learning (IML), Model Interpretability, Model Explainability

1 LECTURER/PROPOSER CONTACT INFORMATION

Name: Dr. Nikos Nikolaou
Affiliation: Centre for Data Intensive Science & Industry, UCL
Address: Room E20, Physics Building, UCL, Gower Street, WC1E 6BT, London, UK
Email: n.nikolaou@ucl.ac.uk
Personal Website: https://nnikolaou.github.io

2 GENERAL COURSE INFORMATION

Course Title: Explainable Machine Learning
Course Category: Introductory Course
Relevant Area: Safe, Explainable and Trustworthy AI (SET)
Course Dates: 22-26 July (there is flexibility on lecture times)

3 COURSE CONTENT INFORMATION

3.1 Motivation

Machine Learning (ML) models are exhibiting impressive performance in a wide range of tasks, rivaling and -in many cases- surpassing that of human-experts or state-of-the-art mechanistic models. As ML models grow in complexity, having established their success, the biggest obstacle in their adoption is now usually their perceived *black box* nature. Using them might give us very accurate predictions, but little-to-no explanation about them. The lack of *transparency* leads to a gap in *trust* and *accountability*.

Explaining the inner workings or the predictions of ML models is especially crucial -for obvious reasons- in critical and high-stakes applications (e.g. medical, legal, security, defence, safety). In these settings *model explainability* is usually an *ethical* -and often a *regulatory* or even *legal*- requirement. But it is also important in scientific applications (e.g. physics, chemistry, biology) where -usually- the end goal is not to simply make accurate predictions about the world, but to *explain* it, to provide us with an *understanding* of how it works. It can even be argued that explaining model predictions is always useful to include as part of an ML pipeline evaluation as it can allow us to identify weaknesses such as presence

of *bias* or instances of *shortcut learning* in the model, which in turn would reveal issues with the data (biases introduced during data collection and/or preprocessing, *information leakage*, *concept drift*). This can inform subsequent data collection, preprocessing and model training & evaluation. It can ultimately inform us of when and how ML our models can fail.

Explainable Machine Learning (XML) a.k.a. *Explainable Artificial Intelligence (XAI)* refers to the family of methods concerned with making the predictions or inner workings of machine learning models more understandable to humans. It covers a wide range of approaches, from methods that construct inherently more transparent models (e.g. *linear models*, *decision trees*, *decision rules*), to *model-specific* methods applicable *post-hoc* (i.e. after model training) to specific families of models (e.g. deep neural networks, differentiable models), to more generally applicable *model-agnostic* methods applicable to any trained black box model. Explainability methods can also vary based on their theoretical underpinnings (from calculus and probability to game theory and combinatorics), the type of explanation we seek (e.g. set of rules, model internals, characteristic datapoints) and whether they explain prediction on individual datapoints (*local explanations*) or more general model behaviour (*global explanations*).

The proposed course will introduce students to the basic principles and tools of explainable ML. In Day 1 we will cover basic notation and terminology, discuss what '*explanations*' are and when we need them, giving successful examples of applications of XAI. Day 2 will cover the most popular methods for obtaining global model explanations (*surrogate modelling*, *PDP-plots*, *M-plots*, *ALE-plots & feature interaction*) and *attributing importance to individual features* (e.g. *LIME*, *SHAP*, *permutation feature importance*). On Day 3 we focus on methods for *obtaining deep neural network internals* (e.g. *layer activations*, *convolutional filters*, *attention maps*) and *local explanations* (various methods of obtaining *pixel attribution* a.k.a. *saliency maps*). Day 4 will focus on *explanations beyond the individual feature level* (*concept-based explanations*) and discuss the use of *datapoints as explanations* as well as the more elaborate topics of *contrastive* and *causal/counterfactual explanations*, providing a broad yet quite deep and unified investigation of the field.

But XAI is not a panacea. In fact there are several *limitations* with the methods we will discuss and several *impossibility results* associated with them. The field of XAI is still in its infancy and a lot of work is needed for it to reach a stage of maturity that meets its stated objectives. It is indicative that an agreed upon definition for even basic terms like '*interpretability*' or '*explainability*' is missing (in the course we will be using them interchangeably). As for the *quantitative* –or even *qualitative– evaluation* of explanations or explainability methods, there are several measures and procedures proposed, often giving contradictory results. Challenges such as *seeking simple explanations for complex models*, *the existence of multiple explanations*, *quantifying uncertainty over explanations* and *tailoring explanations to the intended audience* hinder progress in XAI. It is important to be aware of these limitations both for knowing how XAI can fail (offer little help or even mislead) us and how to advance the field. The last day of the course (Day 5) will focus on these matters, *evaluation of XAI methods*, *limitations* and *future prospects* of the field.

Unfortunately such a course is missing from most academic ML programmes' syllabi and associated summer/winter schools. This course aims to address this gap and equip students with the basic tools to understand model predictions, their uses and limitations as well as the limitations and challenges of XAI as a field and promising future directions. Upon completion students will have a good theoretical and practical understanding of the area.

3.2 Description & Tentative Outline

The course will run on the week of 22-26 of July (times are flexible) and will follow a lecture format. The lectures will be reasonably interactive and they will be accompanied by a relevant Jupyter notebook featuring a code tutorial in Python of the main methods presented. Course slides & Jupyter notebooks will be made available to students. A tentative schedule is given below (not finalized):

- *Day 1 (Monday 22 July) – Introduction to Explainable Machine Learning*
  ML Overview, notation & terminology – ML models as black boxes – Why explain model predictions? – XAI success stories – What makes a good explanation? – Taxonomy of explainability methods – Inherently interpretable models (decision trees, decision rules, linear models, GAMs) – Basic principles of explainability methods.

- *Day 2 (Tuesday 23 July) – General Feature Attribution Methods*
  Global surrogate modelling – Sensitivity analysis (perturbation, occlusion & permutation feature

importance) - Local surrogate modelling (LIME & scoped rules/anchors, Shapely values & SHAP – Other global model-agnostic methods (PDP Plots, M-plots & ALE plots, Feature interaction).

- *Day 3 (Wednesday 24 July) – Explaining Deep Learning Models*
  Deep learning overview – Visualizing model internals (e.g. layer activations, convolutional filters, attention maps) –Saliency maps (Vanilla Gradient, Gradient x Input, Integrated Gradients, Guided Backpropagation, Grad-CAM, Guided Grad-CAM, GradCAM++, SmoothGrad, FusionGrad...) – Are these just glorified edge detectors? – Examples mainly from imaging but also on text, time series & graph data.

- *Day 4 (Thursday 25 July) – Beyond Simple Feature Attribution*
  Detecting Concepts (TCAV, ACE, CBM, CW, multimodal models, ..) – Contrastive explanations – Counterfactual explanations – Examples as explanations (influential instances, prototypes, adversarial examples, counterfactual examples).

- *Day 5 (Friday 26 July) – The Limits of Explainability & Future Prospects*
  Bad examples of XAI – Accuracy vs. explainability – Do we always need to explain model predictions? – Simple explanations in a complex world – Explainable to whom? – Novelty vs. conforming to prior knowledge – The world through the lens of the model – The problem of multiple explanations ('Rashomon effect') – Impossibility results in XAI – Measuring the effects of uncertainty – Evaluating model interpretability (qualitatively & using quantitative metrics, e.g. faithfulness/fidelity, robustness/stability/sensitivity, randomness, complexity, fairness, localization) – Making complex models more inherently interpretable (e.g. incorporating physics / generative / causal constraints into model training) – Automating explainability – Ethical considerations.

3.3 Expected level & prerequisites

The course requires a solid understanding of ML principles & terminology. A basic knowledge of calculus (e.g. concepts of *function*, *partial derivative*, *gradient*) and probability (e.g. concepts of *probability distribution*, *conditional probability*) is desirable. Basic programming knowledge in Python will allow the students to make the most of the accompanying Jupyter Notebook tutorials. The level of understanding textbooks [1-3] below should be sufficient.

3.4 Appropriate references

The course will primarily cover the material from [1][1] & [2]. For explainability methods for deep learning models (Day 3) we will also use material from [3]. Key research articles on the subject of machine learning interpretability & explainability, such as [4-10] (list not exhaustive) will also be used as references.

[1] Molnar, C. (2020). Interpretable machine learning. Lulu.com.
[2] Chollet, F. (2021). Deep learning with Python. Simon and Schuster.
[3] Thampi, A. (2022). Interpretable AI: Building explainable machine learning systems. Simon and Schuster.
[4] Lipton, J. C. (2016). The Mythos of Model Interpretability. 2016 ICML Workshop on Human Interpretability in Machine Learning arXiv preprint arXiv:1606.03490.
[5] Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.
[6] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
[7] Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832.
[8] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5), 206-215.
[9] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. Entropy, 23(1), 18.
[10] Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistic Surveys, 16, 1-85.

[1] Available free online at https://christophm.github.io/interpretable-ml-book/

3.5 Will the course appeal to students outside of the main discipline?

Yes. The course provides an overview of principles, methods, applications, limitations and challenges of explaining the predictions and/or inner workings of machine learning models. As such it is relevant for students interested in advancing the theory of these methods. But it is also relevant to anyone interested in applying ML methods to solve practical problems. XAI methods can allow them to extract new knowledge from their models, to identify biases in their models–or data, to establish trust in their model and –in some cases it is a regulatory requirement. Since ML/AI-based systems are affecting more and more aspects of our life, I would argue that the course might even appeal to a more general audience (i.e. beyond students of theoretical or applied ML). Learning how we can better understand how these systems operate, how to interrogate and audit them, how to inspect their trustworthiness and how to get explanations regarding decisions that affect our lives could be of interest to the general public. After all, under the *EU GDPR* all citizens have the '*right to an explanation*'[2] in such cases (GDPR Article 22). This course can help the students understand what we mean by '*explanations*', how we obtain them, how they can help us, how they can mislead us and how they can be improved.

4 DETAILED LECTURER/PROPOSER INFORMATION

4.1 Short CV of the lecturer/proposer

Dr. N. Nikolaou is a Lecturer (Assistant Professor) at UCL's Centre for Data Intensive Science & Industry (2022- present). His research interests span both the theory of ML (with a focus on *explainable ML*, *uncertainty quantification*, *causal inference* & *resource-efficient ML*) and scientific applications of ML (primarily in *astronomy*, *biology*, *medicine* & *clean energy*). Before his current post, he worked in AstraZeneca Oncology Biometrics R&D as a Senior Data Scientist (2021-22). Prior to this he held a Senior Research Fellow post in the Department of Physics & Astronomy at UCL (2018-21), applying ML methods, with an emphasis on *explainable deep learning*, to exoplanet detection & atmosphere characterization and before that he held an EPSRC Doctoral Prize Fellowship in the Department of Computer Science of the University of Manchester (2016-18), investigating theoretical aspects of ML (*classifier calibration*, *information theory* & *model selection*). He holds a PhD in Computer Science (Machine Learning) from the University of Manchester (2016). His work has received several awards & distinctions: a nomination for the *British Computing Society's (BCS) 2017 'Distinguished Dissertation Award'*, the *2017 'Steve Furber Medal'* & the *2016 'Carol Goble Medal'* by the University of Manchester, a *2016 'Best of Computing, Notable Article'* by the *Association of Computing Machinery (ACM)*, a plenary presentation in the *2016 European Conference on Machine Learning (ECML)*, the *Best Poster Award in the 2015 INIT/AERFAI Machine Learning Summer School*.

4.2 Evidence of relevant teaching experience

Dr. N. Nikolaou has extensive experience communicating ML concepts to audiences of various technical levels in both academia & industry. Since the beginning of his PhD, he has been involved in teaching as a TA (2012-16), invited lecturer (2013-21) or module lead (2022-present). In 2022 he designed and has since been leading the yearly 5-day intensive course *Introduction to Machine Learning* at UCL's Department of Physics & Astronomy (P&A). The target audience consists of MSc & PhD students of UCL P&A with no prior exposure to ML. While at his previous post as Senior Data Scientist at AstraZeneca Oncology (2021-22), he delivered 4 tutorials of the *Deep Teaching* series, aimed at educating non-expert AZ employees on ML-related subjects (*Probability Calibration of ML Classifiers*, *Feature Selection*, *Variational Autoencoders*, & *ML Model Interpretability*). All 4 included a lecture and a hands-on coding component. He has delivered >40 talks –including 24 Invited, and a Plenary Talk at ECML 2016)– to audiences of various technical levels (from non-technical, to world-class expert) in fields ranging from ML & statistics (theory areas) to physics & pharmaceutics/biomedicine (application areas). Audience sizes ranged from $\sim$10 to $\sim$1000 participants (2013-present). He also participated in several outreach activities involving communicating the basic principles and methods of ML and applications in physics & medicine to high-school students, or to audiences consisting of non-ML experts, primarily physicists & medical professionals (2013-present). He has been involved in the supervision of 9 PhD students, 27 MSc students, 1 BSc student summer project & 4 industrial placement group projects involving 12 PhD students, several of whom (4 MSc & 1 PhD) have received prizes for their research projects (2019-present).

[2] https://en.wikipedia.org/wiki/Right_to_explanation