2.8 Interpretability techniques

Practical guidance – cross-domain

Authors: Rhys Ward

One way to provide assurance is to make the ML system being used interpretable. Interpretability may help us to:

- Understand the system retrospectively: to understand, with respect to a harm-causing action or decision, what went wrong, and why
- Understand the system prospectively: to predict, mitigate, and prevent future harm-causing actions or decisions.

In some sense an algorithm is interpretable if we can understand how it works and/or why it makes the decisions that it does make. [17] defines interpretability in the context of ML as 'the ability to explain or to present in understandable terms to a human' but notes that what constitutes an explanation is not well-defined. In practice, the term interpretability is used to refer to a number of distinct concepts. We want to answer the question 'to what extent does machine learning need to be interpretable to provide assurance?'. To answer this question we must decide who needs to understand the system, what they need to understand, what types of explanations are appropriate, and when do these explanations need to be provided.

Types of interpretability

[41] seeks to clarify the myriad different notions of interpretability of ML models in the literature - what interpretability means and why it is important. It is noted that interpretability is not a monolithic concept and relates to a number of distinct ideas. The distinction is often made between methods which are intrinsically transparent and post-hoc methods which attempt to explain a model. We identify the following types of interpretability. A model/system is:

- Transparent if we understand how it works (mechanistically, at some level, for some part of the process). A transparent model is one which is simple enough for humans to understand. We may have transparency at the level of the:
    - Learning algorithm
    - Learned model
    - System logic
    - Parameters or model structures (do they relate to human-understandable concepts?)
- Explainable if we understand why it makes the decisions/predictions that it does make.
    - Global explainability techniques approximate the model with a simpler more transparent one. This simple approximate model is an explanation.

- Local explainability techniques map inputs to outputs and identify important inputs. Other methods locally approximate the model. These methods help us to answer the question 'what were the important factors in this decision?'.

We can also categorise some of the features of these different types of interpretability. Are they faithful representations of the model, or approximations? Do they interpret the whole model (global) or individual decisions (local)? Transparency is an intrinsic property of a model (it is either easy to understand or not, or some degree in between), whereas explainability techniques are post-hoc methods which require some extra effort to implement. Table 1 summarises this.

	Faithful/Approximate	Global/Local	Intrinsic/Post-hoc
Transparency	F	G/L	I
Global Ex.	A	G	PH
Local Ex.	F/A	L	PH

Table 1: Features of different types of interpretability

Interpretability techniques

There is extensive literature surrounding different techniques to interpret or explain ML models or systems. [1] provides a thorough review of current interpretability techniques as summarised in Table 2.

Techniques	References	Intrinsic/Post-hoc	Global/Local	Model-specific/ Model-agnostic
Decision trees	[4] [30] [27] [3] [61]	I	G	SP
Rule lists	[44] [61] [74] [72] [19]	I	G	SP
LIME	[57] [66] [56] [67]	H	L	AG
Shapely explanations	[42]	H	L	AG
Saliency map	[29] [36] [77] [53] [43] [21] [15]	H	L	AG
Activation Maximization	[11] [14]	H	G	AG
Surrogate models	[51] [64] [57]	H	G/L	AG
Partial Dependency Plot (PDP)	[26] [7] [35]	H	G/L	AG
Individual Conditional Expectation (ACI)	[10] [22]	H	L	AG
Rule extraction	[23] [71] [2] [28] [55] [25]	H	G/L	AG
Decomposition	[58] [62] [50]	H	L	AG
Model distillation	[27] [38] [75] [68] [34] [33]	H	G	AG
Sensitive analysis	[13] [12]	H	G/L	AG
Layer-wise Relevence Propagation (LRP)	[49]	H	G/L	AG
Feature Importance	[22] [59] [69]	H	G/L	AG
Prototype and criticism	[8] [60] [16] [37]	H	G/L	AG
Counterfactual explanations	[45]	H	L	AG

Table 2: Summary of interpretability techniques

Table 2 differentiates between local and global explainability i.e. the interpretability of a single decision vs the interpretability of the whole logic of a model. In [41] they also differentiate between intrinsic explainability (e.g. transparency) - simple models which are inherently easy to understand, and post-hoc explainability - methods that analyse the model after training. Post-hoc techniques refer to the global and local techniques described earlier.

Note that each intrinsic technique is also global. This is because this survey considers models which are intrinsically transparent to be transparent globally. In this sense, models cannot be transparent (simple enough for us to understand) for some decisions but not others. The distinction is also made between techniques which are model specific vs model agnostic. A number of other surveys have also been conducted:

- [5] surveys ML methods as they relate to assurance at each stage of the ML life-cycle (Data Management, Model Learning, Model Verification, and Model Deployment).
- [9] surveys interpretable models differentiating between intrinsically explainable and justifiable models/decisions (i.e. transparency vs local explainability).
- [46] surveys different interpretability techniques and compares them on their effectiveness to different user-groups. We will discuss how to evaluate explanations in section 3.
- [18] summarises recent developments in explainable supervised learning.
- [78] reviews recent studies in understanding neural-network representations and learning neural networks with interpretable/disentangled middle-layer representations.
- [48] Seeks to investigate "What makes for a good explanation?" with reference to AI systems and takes a psychological approach. It discusses some explanation methods (e.g. visualisation, text based).
- [24] focuses on explainable methods in deep neural architectures, and briefly highlights review papers from other subfields.
- [47] describes some model-agnostic interpretablity methods, their pros and cons, and how/when to implement them.

Comparing the interpretability of different machine learning models

Table 3 summarises some of the multitude of interpretability techniques in the literature. We classify these by the type of interpretability which they capture (see section 1.1) and by the type of model which they can be used to interpret. Some techniques can be used on multiple models and are referred to as 'Model Agnostic'. Some methods provide some transparency to the system logic without necessarily interpreting the ML model(s) being used (e.g. [20]).

Model Type	Interpretability Type		
	Local Ex.	Global Ex.	Transparency
System Level	[70]		[20]
Model Agnostic	[39] [42] [73]	[32] [40]	
Supervised Learning	[6] [18] [41]	[18] [41]	[18] [41]
Unsupervised Learning	[42] [57]		
Reinforcement Learning (RL)	[31]		
Classifiers	[31] [54] [57]	[40]	
Neural Networks	[6] [24] [78] [65] [76]	[24] [52] [78] [63]	[24]

Table 3: Interpretability techniques for different ML methods

Some of these methods offer very technical "explanations" which would not be suitable for most stakeholders (e.g. doctors, lay-users) and different stakeholders require different types of explanation [46]. When explaining algorithmic decisions the format of the explanation is key.

Summary of approach

1. Define the extent to which the ML system needs to be interpretable and define a set of interpretability requirements (e.g. 'Local decisions can be explained to identify the cause of accidents after they occur') – see guidance on interpretability requirements.
2. Define the types of interpretability needed to meet requirements (e.g. 'local explainability methods implemented to map inputs to outputs and identify important inputs') and consider model selection trade-offs.
3. Implement suitable interpretability techniques (this may result in choosing a transparent model).
4. Ensure explanations are provided in a suitable format for, and are made available to, the audience.
5. Evaluate explanations (see guidance on interpretability evaluation):
    a. Are they suitable for the audience?
    b. Are they faithful to the system process?

References

- [1] Amina Adadi and Mohammed Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: IEEE (2018).
- [2] M. H. Aung et al. "Comparing analytical decision support models through Boolean rule extraction: A case study of ovarian tumour malignancy". In: Int. Symp. Neural Netw. Berlin, Germany: Springer (2007).
- [3] R. D. Gibbons et al. "The CAD-MDD: A computerized adaptive diagnostic screening tool for depression". In:J. Clin. Psychiatry, vol. 74, no. 7, pp. 669–674 (2013).
- [4] V. Schetinin et al. "Confident interpretation of Bayesian decision tree ensembles for clinical applications". In: IEEE Trans. Inf. Technol. Biomed., vol. 11, no. 3, pp. 312–319 (2007).
- [5] Rob Ashmore, Radu Calinescu, and Colin Paterson. "Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges". In: https://arxiv.org/pdf/1905.04223.pdf. May 2019.
- [6] Anand Avati et al. "Improving Palliative Care with Deep Learning". In: arXiv:1711.06402v1 [cs.CY] 17 Nov 2017 (2017).
- [7] R. Berk and J. Bleich. "Statistical procedures for forecasting criminal behavior: A comparative assessment". In: Criminal. Public Policy, vol.12, no. 3, pp. 513–544(2013).
- [8] J. Bien and R. Tibshirani. "Prototype selection for interpretable classification," in: Ann. Appl. Statist., vol. 5, no. 4, pp. 2403–2424 (2011).
- [9] Or Biran and Courtenay Cotton. "Explanation and Justification in Machine Learning: A Survey". In:IJCAI-17 workshop on explainable AI (XAI). Vol. 8. 2017.

- [10] A. Goldstein A. Kapelner J. Bleich and E. Pitkin. “Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation”. In: J. Comput. Graph. Statist., vol. 24, no.1, pp. 44–65 (2015).
- [11] A. Nguyen A. Dosovitskiy J. Yosinski T. Brox and J. Clune. “Synthesizing the preferred inputs for neurons in neural networks via deep generator networks”. In: Adv. Neural Inf. Process. Syst. (NIPS) (2016).
- [12] P. Cortez and M. J. Embrechts. “Opening black box data mining models using sensitivity analysis,” in: IEEE Symp. Comput. Intell. Data Mining (CIDM) (2011).
- [13] P. Cortez and M. J. Embrechts. “Using sensitivity analysis and visualization techniques to open black box data mining models”. In: Inf.Sci., vol. 225, pp. 1–17 (2013).
- [14] A. Courville D. Erhan and Y. Bengio. “Understanding representations learned in deep architectures”. In: Dept. d’Informatique Recherche Operationnelle, Univ. Montreal, Montreal, QC, Canada, Tech. Rep. 1355 (2010).
- [15] P. Dabkowski and Y. Gal. “Real time image saliency for black box classifiers”. In: Adv. Neural Inf. Process. Syst. (2017).
- [16] K. S. Gurumoorthy A. Dhurandhar and G. Cecchi. “ProtoDash: Fast interpretable prototype selection.” In: https://arxiv.org/abs/1707.01212 (2017).
- [17] Finale Doshi-Velez and Been Kim. “Towards A Rigorous Science of Interpretable Machine Learning”. In: arXiv:1702.08608v2 [stat.ML] 2Mar 2017 (2017).
- [18] Filip Karlo Dosilovi, Mario Brci, and Nikica Hlupi. “Explainable Artificial Intelligence: A Survey”. In: MIPRO 2018, May 21-25, 2018, Opatija Croatia (2018).
- [19] D. M. Malioutov K. R. Varshney A. Emad and S. Dash. “Learning interpretable classification rules with boolean compressed sensing”. In: Transparent Data Mining for Big and Small Data. Springer (2017).
- [20] Jeffrey De Fauw et al. “Clinically applicable deep learning for diagnosis and referral in retinal disease”. In: Nature Medicine(2018).
- [21] R. Fong and A. Vedaldi. “Interpretable explanations of black boxes by meaningful perturbation.” In: https://arxiv.org/abs/1704.03296 (2017).
- [22] C. Molnar G. Casalicchio and B. Bischl. “Visualizing the feature importance for black box models.” In: https://arxiv.org/abs/1804.06620 (2018).
- [23] M. van Gerven G. Ras and P. Haselager. “Explanation methods in deep learning: Users, values, concerns and challenges.” In: https://arxiv.org/abs/1803.07517 (2018).
- [24] Leilani H. Gilpin et al. “Explaining Explanations: An Overview of Interpretability of Machine Learning”. In: arXiv:1806.00069v3 [cs.AI]3 Feb 2019 (2019).
- [25] T. GopiKrishna. “Evaluation of rule extraction algorithms”. In: Int. J. Data Mining Knowl. Manage. Process, vol. 4, no. 3, pp. 9–19 (2014).
- [26] D. P. Green and H. L. Kern. “Modeling heterogeneous treatment effects in large-scale experiments using Bayesian additive regression trees”. In: Annu. Summer Meeting Soc. Political Methodol. (2010).
- [27] G. Hooker H. F. Tan and M. T. Wells. “Tree space prototypes: Another look at making tree ensembles interpretable.” In: https://arxiv.org/abs/1611.07115 (2016)
- [28] T. Hailesilassie. “Rule extraction algorithm for deep neural net- works: A review.” In: https://arxiv.org/abs/1610.05267 (2017).

- [29] D. Baehrens T. Schroeter S. Harmeling M. Kawanabe K. Hansen and K.-R. Muller. "How to explain individual classification decisions". In: Mach. Learn. Res., vol. 11, no. 6, pp. 1803–1831 (2010).
- [30] S. Hara and K. Hayashi. "Making tree ensembles interpretable." In:https://arxiv.org/abs/1606.05390 (2016).
- [31] Lisa Anne Hendricks et al. "Generating Visual Explanations". In: arXiv:1603.08507v1 [cs.CV] 28 Mar 2016 (2016).
- [32] Irina Higgins et al. "LEARNING BASIC VISUAL CONCEPTS WITHA CONSTRAINED VARIATIONAL FRAMEWORK". In: ICLR 2017(2017).
- [33] S. Tan R. Caruana G. Hooker and A. Gordo. "Transparent model distillation." In: https://arxiv.org/abs/1801.08640 (2018).
- [34] S. Tan R. Caruana G. Hooker and Y. Lou. "Auditing black-box models using transparent model distillation with side information." In: https://arxiv.org/abs/1710.06169 (2018).
- [35] J. Leathwick J. Elith and T. Hastie. "A working guide to boosted regression trees". In: J. Animal Ecol., vol. 77, no. 4, pp. 802–813 (2008).
- [36] A. Vedaldi K. Simonyan and A. Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: https://arxiv.org/abs/1312.6034 (2013).
- [37] B. Kim, R. Khanna and O. O. Koyejo. "Examples are not enough, learn to criticize! criticism for interpretability," in:29th Conf. Neural Inf. Process. Syst. (NIPS) (2016).
- [38] Z. Che S. Purushotham R. Khemani and Y. Liu. "Distilling knowledge from deep networks with applications to healthcare domain." In: https://arxiv.org/abs/1512.03542 (2015).
- [39] Pang Wei Koh and Percy Liang. "Understanding black-box predictions via influence functions". In: ICML'17 Proceedings of the 34th International Conference on Machine Learning - Volume 70 Pages 1885-1894 (2017).
- [40] Himabindu Lakkaraju et al. "Interpretable Explorable Approximations of Black Box Models". In: arXiv:1707.01154v1 [cs.AI] 4 Jul 2017. 2017.
- [41] Zachary C. Lipton. "The Mythos of Model Interpretability". In: arXiv:1606.03490v3 [cs.LG] (2017).
- [42] S. M. Lundberg and S. I. Lee. "A unified approach t interpreting model predictions". In: Adv. Neural Inf. Process. Syst. (2017).
- [43] A. Taly M. Sundararajan and Q. Yan. "Axiomatic attribution for deep networks." In: https://arxiv.org/abs/1703.01365 (2017).
- [44] B. Letham C. Rudin T. H. McCormick and D. Madigan. "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model". In: Ann. Appl. Statist., vol. 9, no. 3, pp. 1350–1371 (2015).
- [45] S. Wachter B. Mittelstadt and C. Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." In: https://arxiv.org/abs/1711.00399 (2017).
- [46] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. "A Survey of Evaluation Methods and Measures for Interpretable Machine Learning". In: arXiv:1811.11839v2 [cs.HC] 4 Dec 2018 (2018).
- [47] Christoph Molnar. "Interpretable Machine Learning A Guide for Making Black Box Models Explainable". (2019).

- [48] Shane T. Mueller. “Explanation in Human-AI Systems: A LiteratureMeta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI”. In: DARPA XAI Literature Review (2019).
- [49] S. Bach A. Binder G. Montavon F. Klauschen K.-R. Muller and W. Samek. “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” in: PLoS ONE, vol. 10, no. 7,p. e0130140 (2015).
- [50] S. Bach A. Binder K.-R. Muller and W. Samek. “Controlling explanatory heatmap resolution and semantics via decomposition depth”. In: IEEE Int. Conf. Image Process. (ICIP) (2016).
- [51] C. Kim O. Bastani and H. Bastani. “Interpretability via model extraction”. In: https://arxiv.org/abs/1706.09773 (2017).
- [52] Chris Olah, Ludwig Schubert, and Alexander Mordvintsev. “Feature Visualization - How neural networks build up their understanding of images”. In: https://distill.pub/2017/feature-visualization (2017).
- [53] B. Zhou A. Khosla A. Lapedriza A. Oliva and O. Torralba. “Learning deep features for discriminative localization”. In: IEEE Conf. Comput. Vis. Pattern Recognit. (2016)
- [54] Clemens Otte. “Safe and Interpretable Machine Learning A Methodological Review”. In: Computational Intelligence in Intelligent Data Analysis, SCI 445, pp. 111–122. (2013).
- [55] J. Diederich R. Andrews and A. B. Tickle. “Survey and critique of techniques for extracting rules from trained artificial neural networks”. In: Knowl.-Based Syst., vol. 8, no. 6, pp. 373–389 (1995).
- [56] Pedreschi F. Turini R. Guidotti A. Monreal S. Ruggieri D and F.Giannotti. “Local rule-based explanations of black box decision systems”. In: https://arxiv.org/abs/1805.10820 (2018).
- [57] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ““Why Should I Trust You?” Explaining the Predictions of Any Classifier”. In: https://arxiv.org/abs/1602.04938 (2016).
- [58] M. Robnik-ˇSikonja and I. Kononenko. “Explaining classifications forindividual instances”. In:IEEE Trans. Knowl. Data Eng., vol. 20, no.5, pp. 589–600 (2008).
- [59] A. Fisher C. Rudin and F. Dominici. “Model class reliance: Variable importance measures for any machine learning model class, from the ‘rashomon’ perspective.” In: https://arxiv.org/abs/1801.01489 (2018).
- [60] B. Kim C. Rudin and J. A. Shah. “The Bayesian case model: A generative approach for case-based reasoning and prototype classification,” in: Adv. Neural Inf. Process. Syst. (2014).
- [61] A. Fernandez S. Garcıa and F. Herrera. “Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems”. In: Appl. Soft Comput., vol. 9, no. 4, pp. 1304–1314 (2009).
- [62] G. Montavon S. Lapuschkin A. Binder W. Samek and K.-R. M uller.“Explaining nonlinear classification decisions with deep taylor decomposition”. In: Pattern Recognit., vol. 65, pp. 211–222 (2017).
- [63] S. Sarkar. “Accuracy and interpretability trade-offs in machine learning applied to safer gambling”. In: CEUR Workshop Proceedings.

- [64] J. J. Thiagarajan B. Kailkhura P. Sattigeri and K. N. Ramamurthy. “TreeView: Peeking into deep neural networks via feature-space partitioning.” In: https://arxiv.org/abs/1611.07429 (2016).
- [65] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep In-side Convolutional Networks: Visualising Image Classification Modelsand Saliency Maps”. In:https://arxiv.org/abs/1312.6034 (2013).
- [66] M. T. Ribeiro S. Singh and C. Guestrin. “Anchors: High-precision model-agnostic explanations”. In: AAAI Conf. Artif. Intell. (2018).
- [67] S. Mishra, B.L. Stu and S. Dixon. “Local interpretable model-agnostic explanations for music content analysis”. In:ISMIR (2017).
- [68] S. Tan. “Interpretable approaches to detect bias in black-box models”. In: AAAI/ACM Conf. AI Ethics Soc. (2017).
- [69] J. Lei, M. G’Sell, A. Rinaldo, R. J. Tibshirani and L. Wasserman. “Distribution-free predictive inference for regression,” in: http://www.stat.cmu.edu/ryan-tibs/papers/conformal.pdf (2018).
- [70] Nenad Tomasev, Xavier Glorot, and Shakir Mohamed. “A clinically applicable approach to continuous prediction of future acute kidney injury”. In: (2019).
- [71] R. K˙onig U. Johansson and I. Niklasson. “The truth is in there—Rule extraction from opaque models using genetic programming”. In: FLAIRSConf. (2004).
- [72] G. Su D. Wei K. R. Varshney and D. M. Malioutov. “Interpretable two-level Boolean rule learning for classification.” In: https://arxiv.org/abs/1511.07361 (2015).
- [73] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR”. In: CoRRabs/1711.00399 (2017). arXiv:1711.00399.url:http://arxiv.org/abs/1711.00399.
- [74] F. Wang and C. Rudin. “Falling rule lists”. In: Proc. 18th Int. Conf. Artif. Intell. Statist. (AISTATS) (2015).
- [75] K. Xu D. H. Park D. H. Yi and C. Sutton. “Interpreting deep classifier by visual distillation of dark knowledge.” In: https://arxiv.org/abs/1803.04042(2018).
- [76] Tom Zahavy, Nir Ben Zrihem, and Shie Mannor. “Graying the blackbox: Understanding DQNs”. In: https://arxiv.org/abs/1602.02658 (2016).
- [77] M. D. Zeiler and R. Fergus. “Visualizing and understanding convolutional networks”. In: Eur. Conf. Comput. Vis. Zurich, Switzerland: Springer (2014).
- [78] Quan-shi Zhang and Song-chun Zhu. “Visual interpretability for deep learning: a survey”. In: Frontiers of Information Technology Electronic Engineering (2018).