Explaining Explanation Methods

Riccardo Guidotti[0000−0002−2827−7613]

University of Pisa, Italy
riccardo.guidotti@unipi.it

Abstract. The most effective Artificial Intelligence (AI) systems exploit complex machine learning models to fulfill their tasks due to their high performance. Unfortunately, the most effective machine learning models use for their decision processes a logic not understandable from humans that makes them real black-box models. The lack of transparency on how AI systems make decisions is a clear limitation in their adoption in safety-critical and socially sensitive contexts. Consequently, since the applications in which AI are employed are various, research in eXplainable AI (XAI) has recently caught much attention, with specific distinct requirements for different types of explanations for different users. In this paper, we briefly present the existing explanation problems, the main strategies adopted to solve them, and the desiderata for XAI methods. Finally, the most common types of explanations are illustrated with references to state-of-the-art explanation methods able to retrieve them.

1 Introduction

Nowadays, Artificial Intelligence is one of the most important scientific and technological areas, with a huge socio-economic impact and a pervasive adoption in every field of the modern society. High-profile applications such as autonomous vehicles, medical diagnosis, spam filtering, image recognition, and voice assistants are based on Artificial Intelligence (AI) systems. Modern AI is mainly based on Machine Learning models that allow AI systems to reach impressive performance in emulating human behavior. The most effective ML models are *black-box* models [18], i.e., obscure decision-making or predictive methods that "hide" the logic of their internal decision processes to humans, either because not human-understandable or because not directly accessible. Examples of black-box models include Neural Networks and Deep Neural Networks, SVMs, Ensemble classifiers such as Random Forest, but also compositions of expert systems, data mining, and hard-coded software. The choice for the adoption of these obscure models is driven by the high performance in terms of accuracy [36]. As a consequence, the last decade has witnessed the rise of a black-box society [27].

The lack of explanations of how these black-box models make decisions is a restriction for their adoption in safety-critical contexts and socially sensitive domains such as healthcare or law. Moreover, the problem is not only for lack of transparency but also for possible biases inherited by black-box models from artifacts and preconceptions hidden in the training data of the ML algorithms. Predictive ML models learned on biased datasets may inherit such biases, possibly leading to unfair and wrong decisions. Consequences of biased misclassifications can damage decision-makers and put certain societal groups at risk [9, 28, 39] For instance, the AI software used by Amazon to determine the areas of the US to which Amazon would offer free same-day delivery, unintentionally restricted minority neighborhoods from participating in the program (often when every surrounding neighborhood was allowed)[1]. Another example is relative to *propublica.org*. Their journalists have shown that the COMPAS score, a predictive model for the "risk of crime recidivism" (proprietary secret of Northpointe), has a strong ethnic bias. Indeed, according to this score, a black who did not re-offend was classified as "high risk" twice as much as whites who did not re-offend. On the other hand, white repeat offenders were classified as "low risk" twice as much as black repeat offenders[2]. In [9] is shown that the neural network used to train the English language words was encoding biases towards gender and stereotypes. The authors show that for the analogy "Man is to computer programmer as woman is to X", the variable X was replaced by "homemaker" by the neural network. Consequently, the research in eXplainable AI (XAI) and on the study of explanation methods for obscure ML models has recently caught much attention [1, 5, 18, 26, 40].

In addition, an innovative aspect of the *General Data Protection Regulation (GDPR)* promulgated by the European Parliament, which has become law in May 2018, are the clauses on automated decision-making. The GDPR, for the first time, introduces, to some extent, a *right of explanation* for all individuals to obtain "meaningful explanations of the logic involved" when automated decision making takes place. Despite conflicting opinions among legal scholars regarding the real scope of these clauses [15, 24, 37], there is a joint agreement on the need for the implementation of such a principle is imperative and that it represents today a huge open scientific challenge. However, without technology capable of explaining the logic of black boxes, the right to explanation will remain a "dead letter". How can companies trust their AI services without understanding and validating the underlying rationale of their ML components? Furthermore, in turn, how can users trust AI services? It will be impossible to increase the trust of people in AI without explaining the rationale followed by these models. These are the reasons why explanation is now at the heart of responsible, open data science across multiple industry sectors and scientific disciplines.

2 Explanation Methods

A black-box predictor is a ML obscure model, whose internals are unknown to the observer, or they are known but uninterpretable by humans [18]. Therefore,

[1] http://www.techinsider.io/how-algorithms-can-be-racist-2016-4

[2] http://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

in ML to *interpret* means to give or provide the meaning or to *explain* in understandable terms the predictive process of a model to a human [5, 13]. It is assumed that the concepts composing an explanation are self-contained and do not need further explanations [18]. The most widely used approach to explain black-box models and return interpretations is a sort of *reverse engineering*: the explanation is learned by observing the changes in the black-box output by varying the input. A set of *dimensions* are identified to analyze ML interpretability, and explanation methods and, in turn, reflect on existing types of explanations.

Explanation Problems. In the literature, we recognize two types of problems: black box explanation and explanation by design [18]. The *black-box explanation* idea is to couple a ML black-box model with an explanation method able to interpret the black-box decisions. The underlying strategy is to maintain the high performance of the obscure model and to use an explanation method to retrieve the explanations [12, 23, 29]. The explanation methods generally try to approximate the black-box behavior with an interpretable predictor, also named *surrogate* model. This kind of approach is the one more addressed nowadays in the XAI research field. On the other hand, the *explanation by design* consists of directly designing a transparent model that is interpretable by design and aims at replacing the obscure ML model with the new transparent one [32, 33].

In the literature, there are various models recognized to be interpretable. Examples are *decision tree*, *decision rules*, and *linear models* [14]. These models are considered easily understandable and interpretable for humans. They sacrifice performance for interpretability. Besides, most of them cannot be applied to data types such as images or text, but only on tabular data.

Explanation Targets and Strategy. We recognize global and local explanation methods depending on the target of the explanation. A *global* explanation consists in providing an explanation that allows understanding the whole logic of a black-box model and interpreting any possible decision. Global explanations are difficult to achieve, and in the literature are provided only for tabular data. On the other hand, a *local* explanation consists in retrieving the reasons for the prediction returned by a black-box model for a specific case. While for a global explanation, the interpretable surrogate approximates the whole black-box, for a local explanation, the interpretable surrogate model is used to approximate the black-box behavior only in a "neighborhood" of the instance analyzed. The idea is that, in such a neighborhood, it is easier to explain the decision boundary [29].

In addition, we distinguish between model-specific and model-agnostic explanation method depending on the strategy adopted. An explanation method is *model-specific*, or not generalizable [25], if it can be used to interpret only particular types of black-box models. If an explanation method is designed to interpret a Random Forest [36] and internally use a distance between trees, such a method cannot be used to explain the predictions of a Neural Network. On the other hand, a generalizable or *model-agnostic* explanation method can be used independently from the black-box model being explained because the internal characteristics of the black-box are not exploited to retrieve the explanation [29].

Desiderata of Explainable Methods. A set of desiderata should be considered when designing and using explanation methods [14]. The *interpretability* aspect should measure to what extent a given explanation is human-understandable. Interpretability is generally evaluated with the *complexity* of the interpretable surrogate model. For example, the complexity of a rule can be measured with the number of clauses in the condition, for linear models with the number of non-zero weights, while for decision trees with the depth of the tree. The performance of the interpretable surrogate model form which explanations are extracted is generally called *fidelity* and measures to which extent it accurately imitates the black-box prediction. The fidelity is practically measured in terms of Accuracy score, F1-score, etc. [36] with respect to the prediction of the black-box model. Moreover, an interpretable model should satisfy guarantee *fairness* by protecting minorities against discrimination [31], and *privacy* by not revealing sensitive information [2]. Also, an explanation methods must return *robust* and *stable* explanations: similar instances should have similar explanation for a given black-box model [19]. In addition, since the meaningfulness of an explanation depends on the stakeholder [7], the explanation returned must *consider the user background*: common users require simple clarifications, while domain experts can be able to understand complex explanations. Finally, the time that a user is allowed to spend on understanding an explanation is another crucial aspect. In contexts where the decision time is not a constraint, one might prefer a more exhaustive explanation, while when the user needs to quickly make a decision, it is preferable to have an explanation "easy to read". Thus an explanation method must *consider time limitations.*

Types of Explanations. Research on XAI is producing various alternatives. Explanation methods differ one from another depending on the type of explanation returned. In the following, we illustrate the most used types of explanations and highlights how explanation methods build them.

- ***List of Rules.*** An explanation returned in the form of a list of rules implies that rules are read one after the other, and the first rule for which the conditions are verified is used for prediction. Rules are in form of *if-then* rule: *if conditions, then consequent* the *consequent* corresponds to the prediction, while the *conditions* explain the *factual reasons* for the consequent. The *CORELS* method [3] is a transparent by design method able to build a list of rules with the aim of globally replacing the black-box model. A compact set of rules is returned by the transparent predictive method proposed in [21].
- ***Single Tree Approximation.*** The black-box predictor is approximated with a decision tree that represents all the possible decisions. The *TREPAN* explanation method [12] allows to globally explore a Neural Network through a tree structure that, starting from a root, shows for every path the conditions driving the decision process. *TREPAN* retrieves the decision tree by maximizing a gain ratio [36] calculated on the fidelity with respect to the predictions of an obscure Neural Network.
- ***Rule-based Explanation.*** A single if-then rule is used for local explanations. The *conditions* of the rule explain the *factual reasons* for the pre-

diction. The *LORE* explanation method [17] builds a local decision tree in the neighborhood of the instance analyzed, and then extracts from the tree a single rule revealing the reasons for the decision on the specific instance. The *ANCHOR* method [30] returns if then rules called anchors. An anchor contains a set of attributes with the values which are fundamental for obtaining a certain prediction.

- ***Features Importance.*** A feature importance-based local explanation consists of attributes equipped with positive and negative values. The explanation consists of both the sign and the magnitude of the contribution of the attributes for a specific prediction. If the value is positive, then it contributes by increasing the model's output, if the sign is negative, it decreases the output of the model. *LIME* [29] adopts a linear model as the interpretable local surrogate and returns the importance of the features as an explanation exploiting the regression's coefficients. *SHAP* [23] provides the local unique additive feature importance for a specific record exploiting shapely values.
- ***Saliency Maps.*** In image processing, typical explanations consist of *saliency maps*, i.e., images that show the positive (or negative) contribution of each pixel to the black-box prediction. Saliency maps are built for locally explaining DNN models by gradient [34, 35] and perturbation-based [6] attribution methods. These explanation methods assign a score to each pixel such that it is maximized the probability of returning the same answer without considering irrelevant pixels. Under appropriate image transformations that exploit the concept of "superpixels" also methods such as LORE and LIME can be employed to explain black-box working on images.
- ***Prototype-based Explanations.*** An explanation based on *prototypes* returns specimens similar to the instance analyzed, which makes clear the reasons for the prediction. Prototype-based explanations can refer to any type of data. In [11, 22], image prototypes are used as the foundation of the concept for interpretability [8]. In [20] is discussed the concept of *counter-prototypes* called criticisms for tabular data, i.e., prototype showing what should be different to obtain another decision. Exemplar and *counter-exemplars* synthetic images are generated by the *ABELE* explanation method [16] to augment the interpretability of local saliency maps.
- ***Counterfactual Explanations.*** A *counterfactual* explanation shows what should have been different to change the prediction of the black-box model. Counterfactuals help people in reasoning on the cause-effect relations between observed features and classification outcomes [4, 10] and reveal what should change in a given instance to obtain a different prediction [37]. The explanation method proposed in [38] returns counterfactual explanations that describe the smallest change that can be made to a given instance to obtain a certain outcome by solving an optimization problem. The aforementioned LORE [17], besides a factual explanation rule, also provides a set of *counterfactual rules* extracted from the local decision tree, while ABELE [16] returns synthetically generated *counter-exemplar images.*

3 Conclusion

AI systems based on obscure ML models cannot be the long term solution for any real application, especially those involving humans with the final predictions. Research on XAI has strong ethical motivations aimed at empowering users against undesired, possibly illegal, effects of black-box automated decision-making systems. Different types of explanations, and different explanation methods, permits to retrieve the logic of machines, which can be completely different from the logic of humans and resolve unexpected bugs and issues.

However, despite recent developments on XAI some questions remain open. Are the existing explanation methods useful for the realization of the right of explanation declared in the GDPR? Can the actual explanation methods effectively be exploited by business companies for the industrial development of explainable AI services and products? Are explanation methods able to reveal forms of discrimination towards vulnerable social groups, and are their immune from other algorithmic bias and artifacts in the data? Only when these questions will have a positive answer, the research on explanation methods would have reached a satisfactory level.

Acknowledgment

This work is partially supported by the European Community H2020 programme under the funding schemes H2020-INFRAIA-2019-1: Res. Infr. G.A. 871042 *SoBigData++*, (sobigdata.eu), G.A. 952026 *Humane AI-Net*, (humane-ai.eu).

References

1. A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.
2. Y. A. A. S. Aldeen, M. Salleh, and M. A. Razzaque. A comprehensive review on privacy preserving data mining. *SpringerPlus*, 4(1):694, 2015.
3. E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin. Learning certifiably optimal rule lists. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 35–44. ACM, 2017.
4. A. Apicella, F. Isgrò, R. Prevete, and G. Tamburrini. Contrastive explanations to classification systems using sparse dictionaries. In *International Conference on Image Analysis and Processing*, pages 207–218. Springer, 2019.
5. A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.
6. S. Bach, A. Binder, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.
7. U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, et al. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 648–657, 2020.
8. J. Bien and R. Tibshirani. Prototype selection for interpretable classification. *The Annals of Applied Statistics*, 5(4):2403–2424, 2011.

9. T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.
10. R. M. Byrne. Counterfactuals in explainable artificial intelligence (xai): evidence from human reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6276–6282, 2019.
11. C. Chen, O. Li, A. Barnett, J. Su, and C. Rudin. This looks like that: deep learning for interpretable image recognition. *arXiv:1806.10574*, 2018.
12. M. Craven and J. W. Shavlik. Extracting tree-structured representations of trained networks. In *Advances in neural information processing systems*, pages 24–30, 1996.
13. F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
14. A. A. Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1):1–10, 2014.
15. B. Goodman and S. Flaxman. Eu regulations on algorithmic decision-making and a "right to explanation". In *ICML workshop on human interpretability in machine learning (WHI 2016), New York, NY. http://arxiv. org/abs/1606.08813 v1*, 2016.
16. R. Guidotti, A. Monreale, et al. Black box explanation by learning image exemplars in the latent feature space. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 189–205. Springer, 2019.
17. R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, 2019.
18. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.
19. R. Guidotti and S. Ruggieri. On the stability of interpretable models. In *2019 International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2019.
20. B. Kim, O. O. Koyejo, and R. Khanna. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances In Neural Information Processing Systems*, pages 2280–2288, 2016.
21. H. Lakkaraju et al. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1675–1684. ACM, 2016.
22. O. Li, H. Liu, C. Chen, and C. Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Thirty-second AAAI conference on artificial intelligence*, 2018.
23. S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.
24. G. Malgieri and G. Comandé. Why a right to legibility of automated decision-making exists in the General Data Protection Regulation. *International Data Privacy Law*, 7(4):243–265, 2017.
25. D. Martens, B. Baesens, T. Van Gestel, and J. Vanthienen. Comprehensible credit scoring models using rule extraction from support vector machines. *European journal of operational research*, 183(3):1466–1476, 2007.
26. T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.
27. F. Pasquale. *The black box society.* Harvard University Press, 2015.
28. D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri, and F. Turini. Meaningful explanations of black box ai decision systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9780–9784, 2019.

29. M. T. Ribeiro et al. 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*, pages 1135–1144. ACM, 2016.
30. M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2018.
31. A. Romei and S. Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638, 2014.
32. C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *NMI*, 1(5):206–215, 2019.
33. C. Rudin and J. Radin. Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition. *HHDSR*, 1(2), 2019.
34. A. Shrikumar et al. Not just a black box: Learning important features through propagating activation differences. *arXiv:1605.01713*, 2016.
35. K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.
36. P.-N. Tan et al. *Introduction to data mining*. Pearson Education India, 2006.
37. S. Wachter, B. Mittelstadt, and L. Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017.
38. S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *HJLT*, 31:841, 2017.
39. Y. Wang and M. Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *JPSP*, 114(2):246, 2018.
40. Y. Zhang and X. Chen. Explainable recommendation: A survey and new perspectives. *arXiv preprint arXiv:1804.11192*, 2018.