Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges*

Christoph Molnar[1][0000−0003−2331−868X], Giuseppe Casalicchio[1][0000−0001−5324−5966], and Bernd Bischl[1][0000−0001−6002−6980]

Department of Statistics, LMU Munich
Ludwigstr. 33, 80539 Munich, Germany
`christoph.molnar@stat.uni-muenchen.de`

Abstract. We present a brief history of the field of interpretable machine learning (IML), give an overview of state-of-the-art interpretation methods and discuss challenges. Research in IML has boomed in recent years. As young as the field is, it has over 200 years old roots in regression modeling and rule-based machine learning, starting in the 1960s. Recently, many new IML methods have been proposed, many of them model-agnostic, but also interpretation techniques specific to deep learning and tree-based ensembles. IML methods either directly analyze model components, study sensitivity to input perturbations, or analyze local or global surrogate approximations of the ML model. The field approaches a state of readiness and stability, with many methods not only proposed in research, but also implemented in open-source software. But many important challenges remain for IML, such as dealing with dependent features, causal interpretation, and uncertainty estimation, which need to be resolved for its successful application to scientific problems. A further challenge is a missing rigorous definition of interpretability, which is accepted by the community. To address the challenges and advance the field, we urge to recall our roots of interpretable, data-driven modeling in statistics and (rule-based) ML, but also to consider other areas such as sensitivity analysis, causal inference, and the social sciences.

Keywords: Interpretable Machine Learning · Explainable Artificial Intelligence

1   Introduction

Interpretability is often a deciding factor when a machine learning (ML) model is used in a product, a decision process, or in research. Interpretable machine learning (IML)[1] methods can be used to discover knowledge, to debug or justify

---

[1] Sometimes the term Explainable AI is used.

the model and its predictions, and to control and improve the model [1]. In this paper, we take a look at the historical building blocks of IML and give an overview of methods to interpret models. We argue that IML has reached a state of readiness, but some challenges remain.

2   A Brief History of IML

A lot of IML research happened in the last couple of years. But learning interpretable models from data has a much longer tradition. Linear regression models were used by Gauss, Legendre, and Quetelet [109,64,37,90] as early as the beginning of the 19th century and have since then grown into a vast array of regression analysis tools [115,98], for example, generalized additive models [45] and elastic net [132]. The philosophy behind these statistical models is usually to make certain distributional assumptions or to restrict the model complexity beforehand and thereby imposing intrinsic interpretability of the model.

In ML, a slightly different modeling approach is pursued. Instead of restricting the model complexity beforehand, ML algorithms usually follow a non-linear, non-parametric approach, where model complexity is controlled through one or more hyperparameters and selected via cross-validation. This flexibility often results in less interpretable models with good predictive performance. A lot of ML research began in the second half of the 20th century with research on, for example, support vector machines in 1974 [119], early important work on neural networks in the 1960s [100], and boosting in 1990 [99]. Rule-based ML, which covers decision rules and decision trees, has been an active research area since the middle of the 20th century [35].

While ML algorithms usually focus on predictive performance, work on interpretability in ML – although underexplored – has existed for many years. The built-in feature importance measure of random forests [13] was one of the important IML milestones.[2] In the 2010s came the deep learning hype, after a deep neural network won the ImageNet challenge. A few years after that, the IML field really took off (around 2015), judging by the frequency of the search terms "Interpretable Machine Learning" and "Explainable AI" on Google (Figure 1, right) and papers published with these terms (Figure 1, left). Since then, many model-agnostic explanation methods have been introduced, which work for different types of ML models. But also model-specific explanation methods have been developed, for example, to interpret deep neural networks or tree ensembles. Regression analysis and rule-based ML remain important and active research areas to this day and are blending together (e.g., model-based trees [128], RuleFit [33]). Many extensions of the linear regression model exist [45,25,38] and new extensions are proposed until today [26,14,27,117]. Rule-based ML also remains an active area of research (for example, [123,66,52]). Both regression models and

---

[2] The random forest paper has been cited over 60,000 times (Google Scholar; September 2020) and there are many papers improving the importance measure ([110,111,44,56]) which are also cited frequently.
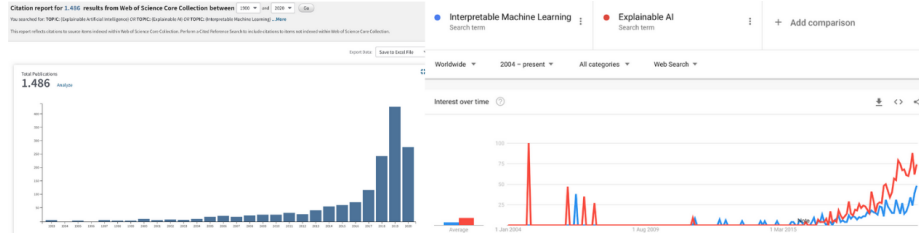
Fig. 1. Left: Citation count for research articles with keywords "Interpretable Machine Learning" or "Explainable AI" on Web of Science (accessed August 10, 2020). Right: Google search trends for "Interpretable Machine Learning" and "Explainable AI" (accessed August 10, 2020).

rule-based ML serve as stand-alone ML algorithms, but also as building blocks for many IML approaches.

3 Today

IML has reached a first state of readiness. Research-wise, the field is maturing in terms of methods surveys [75,41,120,96,1,6,23,15], further consolidation of terms and knowledge [42,22,82,97,88,17], and work about defining interpretability or evaluation of IML methods [74,73,95,49]. We have a better understanding of weaknesses of IML methods in general [75,79], but also specifically for methods such as permutation feature importance [51,110,7,111], Shapley values [57,113], counterfactual explanations [63], partial dependence plots [51,50,7] and saliency maps [2]. Open source software with implementations of various IML methods is available, for example, *iml* [76] and *DALEX* [11] for R [91] and *Alibi* [58] and *InterpretML* [83] for Python. Regulation such as GDPR and the need for ML trustability, transparency and fairness have sparked a discussion around further needs of interpretability [122]. IML has also arrived in industry [36], there are startups that focus on ML interpretability and also big tech companies offer software [126,8,43].

4 IML Methods

We distinguish IML methods by whether they analyze model components, model sensitivity[3], or surrogate models, illustrated in Figure 4.[4]

---

[3] Not to be confused with the research field of sensitivity analysis, which studies the uncertainty of outputs in mathematical models and systems. There are methodological overlaps (e.g., Shapley values), but also differences in methods and how input data distributions are handled.

[4] Some surveys distinguish between *ante-hoc (or transparent design, white-box models, inherently interpretable model)* and *post-hoc* IML method, depending on whether
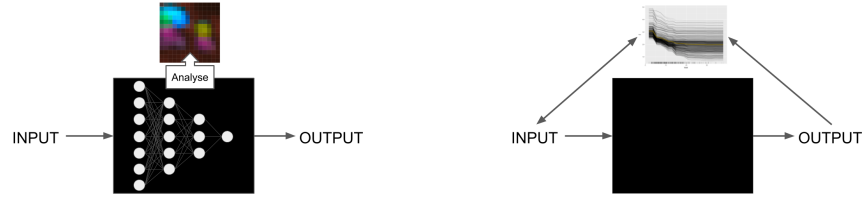
Fig. 2. Some IML approaches work by assigning meaning to individual model components (left), some by analyzing the model predictions for perturbations of the data (right). The surrogate approach, a mixture of the two other approaches, approximates the ML model using (perturbed) data and then analyzes the components of the interpretable surrogate model.

4.1   Analyzing Components of Interpretable Models

In order to analyze components of a model, it needs to be decomposable into parts that we can interpret individually. However, it is not necessarily required that the user understands the model in its entirety (simulatability [82]). Component analysis is always model-specific, because it is tied to the structure of the model.

Inherently interpretable models are models with (learned) structures and (learned) parameters which can be assigned a certain interpretation. In this context, linear regression models, decision trees and decision rules are considered to be interpretable [30,54]. Linear regression models can be interpreted by analyzing components: The model structure, a weighted sum of features, allows to interpret the weights as the effects that the features have on the prediction.

Decision trees and other rule-based ML models have a learned structure (e.g.,"IF feature $x_1 > 0$ and feature $x_2 \in \{A, B\}$, THEN predict 0.6"). We can interpret the learned structure to trace how the model makes predictions.

This only works up to a certain point in high-dimensional scenarios. Linear regression models with hundreds of features and complex interaction terms or deep decision trees are not that interpretable anymore. Some approaches aim to reduce the parts to be interpreted. For example, LASSO [98,115] shrinks the coefficients in a linear model so that many of them become zero, and pruning techniques shorten trees.

4.2   Analyzing Components of More Complex Models

With a bit more effort, we can also analyze components of more complex black-box models. [5] For example, the abstract features learned by a deep convolutional neural network (CNN) can be visualized by finding or generating images that

---

interpretability is considered at model design and training or after training, leaving the (black-box) model unchanged. Another category separates model-agnostic and model-specific methods.

[5] This blurs the line between an "inherently interpretable" and a "black-box" model.

activate a feature map of the CNN [84]. For the random forest, the minimal depth distribution [85,55] and the Gini importance [13] analyze the structure of the trees of the forest and can be used to quantify feature importance. Some approaches aim to make the parts of a model more interpretable with, for example, a monotonicity constraint [106] or a modified loss function for disentangling concepts learned by a convolutional neural network [130].

If an ML algorithm is well understood and frequently used in a community, like random forests in ecology research [19], model component analysis can be the correct tool, but it has the obvious disadvantage that it is tied to that specific model. And it does not combine well with the common model selection approach in ML, where one usually searches over a large class of different ML models via cross-validation.

4.3   Explaining Individual Predictions

Methods that study the sensitivity of an ML model are mostly model-agnostic and work by manipulating input data and analyzing the respective model predictions. These IML methods often treat the ML model as a closed system that receives feature values as an input and produces a prediction as output. We distinguish between local and global explanations.

Local methods explain individual predictions of ML models. Local explanation methods have received much attention and there has been a lot of innovation in the last years. Popular local IML methods are Shapley values [69,112] and counterfactual explanations [122,20,81,116,118]. Counterfactual explanations explain predictions in the form of what-if scenarios, which builds on a rich tradition in philosophy [108]. According to findings in the social sciences [71], counterfactual explanations are "good" explanations because they are contrastive and focus on a few reasons. A different approach originates from collaborative game theory: The Shapley values [104] provide an answer on how to fairly share a payout among the players of a collaborative game. The collaborative game idea can be applied to ML where features (i.e., the players) collaborate to make a prediction (i.e., the payout) [112,69,68].

Some IML methods rely on model-specific knowledge to analyze how changes in the input features change the output. Saliency maps, an interpretation method specific for CNNs, make use of the network gradients to explain individual classifications. The explanations are in the form of heatmaps that show how changing a pixel can change the classification. The saliency map methods differ in how they backpropagate [114,69,80,107,105]. Additionally, model-agnostic versions [95,69,129] exist for analyzing image classifiers.

4.4   Explaining Global Model Behavior

Global model-agnostic explanation methods are used to explain the expected model behavior, i.e., how the model behaves on average for a given dataset. A useful distinction of global explanations are feature importance and feature effect.

Feature importance ranks features based on how relevant they were for the prediction. Permutation feature importance [28,16] is a popular importance measure, originally suggested for random forests [13]. Some importance measures rely on removing features from the training data and retraining the model [65]. An alternative are variance-based measures [40]. See [125] for an overview of importance measures.

The feature effect expresses how a change in a feature changes the predicted outcome. Popular feature effect plots are partial dependence plots [32], individual conditional expectation curves [39], accumulated local effect plots [7], and the functional ANOVA [50]. Analyzing influential data instances, inspired by statistics, provides a different view into the model and describes how influential a data point was for a prediction [59].

4.5   Surrogate Models

Surrogate models[6] are interpretable models designed to "copy" the behavior of the ML model. The surrogate approach treats the ML model as a black-box and only requires the input and output data of the ML model (similar to sensitivity analysis) to train a surrogate ML model. However, the interpretation is based on analyzing components of the interpretable surrogate model. Many IML methods are surrogate model approaches [89,75,72,95,34,10,18,61] and differ, e.g., in the targeted ML model, the data sampling strategy, or the interpretable model that is used. There are also methods for extracting, e.g., decision rules from specific models based on their internal components such as neural network weights [5,9]. LIME [95] is an example of a local surrogate method that explains individual predictions by learning an interpretable model with data in proximity to the data point to be explained. Numerous extensions of LIME exist, which try to fix issues with the original method, extend it to other tasks and data, or analyze its properties [53,93,92,121,47,94,103,12].

5   Challenges

This section presents an incomplete overview of challenges for IML, mostly based on [79].

5.1   Statistical Uncertainty and Inference

Many IML methods such as permutation feature importance or Shapley values provide explanations without quantifying the uncertainty of the explanation. The model itself, but also its explanations, are computed from data and hence are subject to uncertainty. First research is working towards quantifying uncertainty of explanations, for example, for feature importance [124,28,4], layer-wise relevance propagation [24], and Shapley values [127].

---

[6] Surrogate models are related to knowledge distillation and the teacher-student model.

In order to infer meaningful properties of the underlying data generating process, we have to make structural or distributional assumptions. Whether it is a classical statistical model, an ML algorithm or an IML procedure, these assumptions should be clearly stated and we need better diagnostic tools to test them. If we want to prevent statistical testing problems such as p-hacking [48] to reappear in IML, we have to become more rigorous in studying and quantifying the uncertainty of IML methods. For example, most IML methods for feature importance are not adapted for multiple testing, which is a classic mistake in a statistical analysis.

5.2   Causal Interpretation

Ideally, a model should reflect the true causal structure of its underlying phenomena, to enable causal interpretations. Arguably, causal interpretation is usually the goal of modeling if ML is used in science. But most statistical learning procedures reflect mere correlation structures between features and analyze the surface of the data generation process instead of its true inherent structure. Such causal structures would also make models more robust against adversarial attacks [101,29], and more useful when used as a basis for decision making. Unfortunately, predictive performance and causality can be conflicting goals. For example, today's weather directly causes tomorrow's weather, but we might only have access to the feature "wet ground". Using "wet ground" in the prediction model for "tomorrow's weather" is useful as it has information about "today's weather", but we are not allowed to interpret it causally, because the confounder "today's weather" is missing from the ML model. Further research is needed to understand when we are allowed to make causal interpretations of an ML model. First steps have been made for permutation feature importance [60] and Shapley values [70].

5.3   Feature Dependence

Feature dependence introduces problems with attribution and extrapolation. Attribution of importance and effects of features becomes difficult when features are, for example, correlated and therefore share information. Correlated features in random forests are preferred and attributed a higher importance [110,51]. Many sensitivity analysis based methods permute features. When the permuted feature has some dependence with another feature, this association is broken and the resulting data points extrapolate to areas outside the distribution. The ML model was never trained on such combinations and will likely not be confronted with similar data points in an application. Therefore, extrapolation can cause misleading interpretations. There have been attempts to "fix" permutation-based methods, by using a conditional permutation scheme that respects the joint distribution of the data [78,110,28,51]. The change from unconditional to conditional permutation changes the respective interpretation method [78,7], or, in worst case, can break it [57,113,62].

5.4   Definition of Interpretability

A lack of definition for the term "interpretability" is a common critique of the field [67,22]. How can we decide if a new method explains ML models better without a satisfying definition of interpretability? To evaluate the predictive performance of an ML model, we simply compute the prediction error on test data given the groundtruth label. To evaluate the interpretability of that same ML model is more difficult. We do not know what the groundtruth explanation looks like and have no straightforward way to quantify how interpretable a model is or how correct an explanation is. Instead of having one groundtruth explanation, various quantifiable aspects of interpretability are emerging [87,86,77,46,131,3,102,87,21,31].

The two main ways of evaluating interpretability are objective evaluations, which are mathematically quantifiable metrics, and human-centered evaluations, which involve studies with either domain experts or lay persons. Examples of aspects of interpretability are sparsity, interaction strength, fidelity (how well an explanation approximates the ML model), sensitivity to perturbations, and a user's ability to run a model on a given input (simulatability). The challenge ahead remains to establish a best practice on how to evaluate interpretation methods and the explanations they produce. Here, we should also look at the field of human-computer interaction.

5.5   More Challenges Ahead

We focused mainly on the methodological, mathematical challenges in a rather static setting, where a trained ML model and the data are assumed as given and fixed. But ML models are usually not used in a static and isolated way, but are embedded in some process or product, and interact with people. A more dynamic and holistic view of the entire process, from data collection to the final consumption of the explained prediction is needed. This includes thinking how to explain predictions to individuals with diverse knowledge and backgrounds and about the need of interpretability on the level of an institution or society in general. This covers a wide range of fields, such as human-computer interaction, psychology and sociology. To solve the challenges ahead, we believe that the field has to reach out horizontally – to other domains – and vertically – drawing from the rich research in statistics and computer science.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access **6**, 52138–52160 (2018)
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems. pp. 9505–9515 (2018)
3. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Selected papers of Hirotugu Akaike, pp. 199–213. Springer (1998)

4. Altmann, A., Toloşi, L., Sander, O., Lengauer, T.: Permutation importance: a corrected feature importance measure. Bioinformatics **26**(10), 1340–1347 (2010)
5. Andrews, R., Diederich, J., Tickle, A.B.: Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-based systems **8**(6), 373–389 (1995)
6. Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: Results from a systematic literature review. In: 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019. pp. 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems (2019)
7. Apley, D.W., Zhu, J.: Visualizing the effects of predictor variables in black box supervised learning models. arXiv preprint arXiv:1612.08468 (2016)
8. Arya, V., Bellamy, R.K., Chen, P.Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilovic, A., et al.: AI explainability 360: An extensible toolkit for understanding data and machine learning models. Journal of Machine Learning Research **21**(130), 1–6 (2020)
9. Augasta, M.G., Kathirvalavakumar, T.: Rule extraction from neural networks—a comparative study. In: International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012). pp. 404–408. IEEE (2012)
10. Bastani, O., Kim, C., Bastani, H.: Interpreting blackbox models via model extraction. arXiv preprint arXiv:1705.08504 (2017)
11. Biecek, P.: DALEX: explainers for complex predictive models in r. The Journal of Machine Learning Research **19**(1), 3245–3249 (2018)
12. Botari, T., Hvilshøj, F., Izbicki, R., de Carvalho, A.C.: MeLIME: Meaningful local explanation for machine learning models. arXiv preprint arXiv:2009.05818 (2020)
13. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
14. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1721–1730 (2015)
15. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8), 832 (2019)
16. Casalicchio, G., Molnar, C., Bischl, B.: Visualizing the feature importance for black box models. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 655–670. Springer (2018)
17. Chromik, M., Schuessler, M.: A taxonomy for human subject evaluation of blackbox explanations in XAI. In: ExSS-ATEC@ IUI (2020)
18. Craven, M., Shavlik, J.W.: Extracting tree-structured representations of trained networks. In: Advances in neural information processing systems. pp. 24–30 (1996)
19. Cutler, D.R., Edwards Jr, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J.: Random forests for classification in ecology. Ecology **88**(11), 2783–2792 (2007)
20. Dandl, S., Molnar, C., Binder, M., Bischl, B.: Multi-objective counterfactual explanations. arXiv preprint arXiv:2004.11165 (2020)
21. Dhurandhar, A., Iyengar, V., Luss, R., Shanmugam, K.: TIP: typifying the interpretability of procedures. arXiv preprint arXiv:1706.02952 (2017)
22. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
23. Du, M., Liu, N., Hu, X.: Techniques for interpretable machine learning. Communications of the ACM **63**(1), 68–77 (2019)

24. Fabi, K., Schneider, J.: On feature relevance uncertainty: A Monte Carlo dropout sampling approach. arXiv preprint arXiv:2008.01468 (2020)
25. Fahrmeir, L., Tutz, G.: Multivariate statistical modelling based on generalized linear models. Springer Science & Business Media (2013)
26. Fasiolo, M., Nedellec, R., Goude, Y., Wood, S.N.: Scalable visualization methods for modern generalized additive models. Journal of computational and Graphical Statistics **29**(1), 78–86 (2020)
27. Fasiolo, M., Wood, S.N., Zaffran, M., Nedellec, R., Goude, Y.: Fast calibrated additive quantile regression. Journal of the American Statistical Association pp. 1–11 (2020)
28. Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research **20**(177), 1–81 (2019)
29. Freiesleben, T.: Counterfactual explanations & adversarial examples– common grounds, essential differences, and potential transfers. arXiv preprint arXiv:2009.05487 (2020)
30. Freitas, A.A.: Comprehensible classification models: a position paper. ACM SIGKDD explorations newsletter **15**(1), 1–10 (2014)
31. Friedler, S.A., Roy, C.D., Scheidegger, C., Slack, D.: Assessing the local interpretability of machine learning models. arXiv preprint arXiv:1902.03501 (2019)
32. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189–1232 (2001)
33. Friedman, J.H., Popescu, B.E., et al.: Predictive learning via rule ensembles. The Annals of Applied Statistics **2**(3), 916–954 (2008)
34. Frosst, N., Hinton, G.: Distilling a neural network into a soft decision tree. arXiv preprint arXiv:1711.09784 (2017)
35. Fürnkranz, J., Gamberger, D., Lavrač, N.: Foundations of rule learning. Springer Science & Business Media (2012)
36. Gade, K., Geyik, S.C., Kenthapadi, K., Mithal, V., Taly, A.: Explainable AI in industry. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 3203–3204 (2019)
37. Gauss, C.F.: Theoria motus corporum coelestium in sectionibus conicis solem ambientium, vol. 7. Perthes et Besser (1809)
38. Gelman, A., Hill, J.: Data analysis using regression and multilevel/hierarchical models. Cambridge university press (2006)
39. Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics **24**(1), 44–65 (2015)
40. Greenwell, B.M., Boehmke, B.C., McCarthy, A.J.: A simple and effective model-based variable importance measure. arXiv preprint arXiv:1805.04755 (2018)
41. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)
42. Hall, M., Harborne, D., Tomsett, R., Galetic, V., Quintana-Amate, S., Nottle, A., Preece, A.: A systematic method to understand requirements for explainable AI(XAI) systems. In: Proceedings of the IJCAI Workshop on eXplainable Artificial Intelligence (XAI 2019), Macau, China (2019)
43. Hall, P., Gill, N., Kurka, M., Phan, W.: Machine learning interpretability with h2o driverless AI. H2O. ai. URL: http://docs. h2o. ai/driverless-ai/latest-stable/docs/booklets/MLIBooklet. pdf (2017)

44. Hapfelmeier, A., Hothorn, T., Ulm, K., Strobl, C.: A new variable importance measure for random forests with missing data. Statistics and Computing **24**(1), 21–34 (2014)
45. Hastie, T.J., Tibshirani, R.J.: Generalized additive models, vol. 43. CRC press (1990)
46. Hauenstein, S., Wood, S.N., Dormann, C.F.: Computing AIC for black-box models using generalized degrees of freedom: A comparison with cross-validation. Communications in Statistics-Simulation and Computation **47**(5), 1382–1396 (2018)
47. Haunschmid, V., Manilow, E., Widmer, G.: audioLIME: Listenable explanations using source separation. arXiv preprint arXiv:2008.00582 (2020)
48. Head, M.L., Holman, L., Lanfear, R., Kahn, A.T., Jennions, M.D.: The extent and consequences of p-hacking in science. PLoS Biol **13**(3), e1002106 (2015)
49. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608 (2018)
50. Hooker, G.: Generalized functional anova diagnostics for high-dimensional functions of dependent variables. Journal of Computational and Graphical Statistics **16**(3), 709–732 (2007)
51. Hooker, G., Mentch, L.: Please stop permuting features: An explanation and alternatives. arXiv preprint arXiv:1905.03151 (2019)
52. Hothorn, T., Hornik, K., Zeileis, A.: ctree: Conditional inference trees. The Comprehensive R Archive Network **8** (2015)
53. Hu, L., Chen, J., Nair, V.N., Sudjianto, A.: Locally interpretable models and effects based on supervised partitioning (LIME-SUP). arXiv preprint arXiv:1806.00663 (2018)
54. Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., Baesens, B.: An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decision Support Systems **51**(1), 141–154 (2011)
55. Ishwaran, H., Kogalur, U.B., Gorodeski, E.Z., Minn, A.J., Lauer, M.S.: High-dimensional variable selection for survival data. Journal of the American Statistical Association **105**(489), 205–217 (2010)
56. Ishwaran, H., et al.: Variable importance in binary regression trees and forests. Electronic Journal of Statistics **1**, 519–537 (2007)
57. Janzing, D., Minorics, L., Blöbaum, P.: Feature relevance quantification in explainable AI: A causality problem. arXiv preprint arXiv:1910.13413 (2019)
58. Klaise, J., Van Looveren, A., Vacanti, G., Coca, A.: Alibi: Algorithms for monitoring and explaining machine learning models. URL https://github.com/SeldonIO/alibi (2020)
59. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. arXiv preprint arXiv:1703.04730 (2017)
60. König, G., Molnar, C., Bischl, B., Grosse-Wentrup, M.: Relative feature importance. arXiv preprint arXiv:2007.08283 (2020)
61. Krishnan, S., Wu, E.: Palm: Machine learning explanations for iterative debugging. In: Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics. pp. 1–6 (2017)
62. Kumar, I.E., Venkatasubramanian, S., Scheidegger, C., Friedler, S.: Problems with Shapley-value-based explanations as feature importance measures. arXiv preprint arXiv:2002.11097 (2020)
63. Laugel, T., Lesot, M.J., Marsala, C., Renard, X., Detyniecki, M.: The dangers of post-hoc interpretability: Unjustified counterfactual explanations. arXiv preprint arXiv:1907.09294 (2019)

64. Legendre, A.M.: Nouvelles méthodes pour la détermination des orbites des comètes. F. Didot (1805)
65. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L.: Distribution-free predictive inference for regression. Journal of the American Statistical Association **113**(523), 1094–1111 (2018)
66. Letham, B., Rudin, C., McCormick, T.H., Madigan, D., et al.: Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. The Annals of Applied Statistics **9**(3), 1350–1371 (2015)
67. Lipton, Z.C.: The mythos of model interpretability. Queue **16**(3), 31–57 (2018)
68. Lundberg, S.M., Erion, G.G., Lee, S.I.: Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888 (2018)
69. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in neural information processing systems. pp. 4765–4774 (2017)
70. Ma, S., Tourani, R.: Predictive and causal implications of using Shapley value for model interpretation. In: Proceedings of the 2020 KDD Workshop on Causal Discovery. pp. 23–38. PMLR (2020)
71. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)
72. Ming, Y., Qu, H., Bertini, E.: Rulematrix: Visualizing and understanding classifiers with rules. IEEE transactions on visualization and computer graphics **25**(1), 342–352 (2018)
73. Mohseni, S., Ragan, E.D.: A human-grounded evaluation benchmark for local explanations of machine learning. arXiv preprint arXiv:1801.05075 (2018)
74. Mohseni, S., Zarei, N., Ragan, E.D.: A multidisciplinary survey and framework for design and evaluation of explainable AI systems. arXiv pp. arXiv–1811 (2018)
75. Molnar, C.: Interpretable Machine Learning (2019), `https://christophm.github.io/interpretable-ml-book/`
76. Molnar, C., Bischl, B., Casalicchio, G.: iml: An R package for interpretable machine learning. JOSS **3**(26), 786 (2018)
77. Molnar, C., Casalicchio, G., Bischl, B.: Quantifying model complexity via functional decomposition for better post-hoc interpretability. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 193–204. Springer (2019)
78. Molnar, C., König, G., Bischl, B., Casalicchio, G.: Model-agnostic feature importance and effects with dependent features–a conditional subgroup approach. arXiv preprint arXiv:2006.04628 (2020)
79. Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., Casalicchio, G., Grosse-Wentrup, M., Bischl, B.: Pitfalls to avoid when interpreting machine learning models. arXiv preprint arXiv:2007.04131 (2020)
80. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition **65**, 211–222 (2017)
81. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 607–617 (2020)
82. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences **116**(44), 22071–22080 (2019)
83. Nori, H., Jenkins, S., Koch, P., Caruana, R.: Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223 (2019)

84. Olah, C., Mordvintsev, A., Schubert, L.: Feature visualization. Distill (2017). https://doi.org/10.23915/distill.00007, https://distill.pub/2017/feature-visualization

85. Paluszynska, A., Biecek, P., Jiang, Y.: randomForestExplainer: Explaining and Visualizing Random Forests in Terms of Variable Importance (2020), `https://CRAN.R-project.org/package=randomForestExplainer`, r package version 0.10.1

86. Philipp, M., Rusch, T., Hornik, K., Strobl, C.: Measuring the stability of results from supervised statistical learning. Journal of Computational and Graphical Statistics **27**(4), 685–700 (2018)

87. Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Vaughan, J.W., Wallach, H.: Manipulating and measuring model interpretability. arXiv preprint arXiv:1802.07810 (2018)

88. Preece, A., Harborne, D., Braines, D., Tomsett, R., Chakraborty, S.: Stakeholders in explainable AI. arXiv preprint arXiv:1810.00184 (2018)

89. Puri, N., Gupta, P., Agarwal, P., Verma, S., Krishnamurthy, B.: Magix: Model agnostic globally interpretable explanations. arXiv preprint arXiv:1706.07160 (2017)

90. Quetelet, L.A.J.: Recherches sur la population, les naissances, les décès, les prisons, les dépôts de mendicité, etc. dans le royaume des Pays-Bas (1827)

91. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2020), `https://www.R-project.org/`

92. Rabold, J., Deininger, H., Siebers, M., Schmid, U.: Enriching visual with verbal explanations for relational concepts–combining LIME with Aleph. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 180–192. Springer (2019)

93. Rabold, J., Siebers, M., Schmid, U.: Explaining black-box classifiers with ilp–empowering LIME with aleph to approximate non-linear decisions with relational rules. In: International Conference on Inductive Logic Programming. pp. 105–117. Springer (2018)

94. Rahnama, A.H.A., Boström, H.: A study of data and label shift in the LIME framework. arXiv preprint arXiv:1910.14421 (2019)

95. Ribeiro, M.T., Singh, S., Guestrin, C.: " 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. pp. 1135–1144 (2016)

96. Rosenfeld, A., Richardson, A.: Explainability in human–agent systems. Autonomous Agents and Multi-Agent Systems **33**(6), 673–705 (2019)

97. Samek, W., Müller, K.R.: Towards explainable artificial intelligence. In: Explainable AI: interpreting, explaining and visualizing deep learning, pp. 5–22. Springer (2019)

98. Santosa, F., Symes, W.W.: Linear inversion of band-limited reflection seismograms. SIAM Journal on Scientific and Statistical Computing **7**(4), 1307–1330 (1986)

99. Schapire, R.E.: The strength of weak learnability. Machine learning **5**(2), 197–227 (1990)

100. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural networks **61**, 85–117 (2015)

101. Schölkopf, B.: Causality for machine learning. arXiv preprint arXiv:1911.10500 (2019)

102. Schwarz, G., et al.: Estimating the dimension of a model. The annals of statistics **6**(2), 461–464 (1978)

103. Shankaranarayana, S.M., Runje, D.: ALIME: Autoencoder based approach for local interpretability. In: International Conference on Intelligent Data Engineering and Automated Learning. pp. 454–463. Springer (2019)
104. Shapley, L.S.: A value for n-person games. Contributions to the Theory of Games **2**(28), 307–317 (1953)
105. Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not just a black box: Learning important features through propagating activation differences. arXiv preprint arXiv:1605.01713 (2016)
106. Sill, J.: Monotonic networks. In: Advances in neural information processing systems. pp. 661–667 (1998)
107. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
108. Starr, W.: Counterfactuals (2019)
109. Stigler, S.M.: The history of statistics: The measurement of uncertainty before 1900. Harvard University Press (1986)
110. Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A.: Conditional variable importance for random forests. BMC bioinformatics **9**(1), 307 (2008)
111. Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics **8**(1), 25 (2007)
112. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems **41**(3), 647–665 (2014)
113. Sundararajan, M., Najmi, A.: The many Shapley values for model explanation. arXiv preprint arXiv:1908.08474 (2019)
114. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. arXiv preprint arXiv:1703.01365 (2017)
115. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996)
116. Tolomei, G., Silvestri, F., Haines, A., Lalmas, M.: Interpretable predictions of tree-based ensembles via actionable feature tweaking. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 465–474 (2017)
117. Ustun, B., Rudin, C.: Supersparse linear integer models for optimized medical scoring systems. Machine Learning **102**(3), 349–391 (2016)
118. Ustun, B., Spangher, A., Liu, Y.: Actionable recourse in linear classification. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 10–19 (2019)
119. Vapnik, V., Chervonenkis, A.: Theory of pattern recognition (1974)
120. Vilone, G., Longo, L.: Explainable artificial intelligence: a systematic review. arXiv preprint arXiv:2006.00093 (2020)
121. Visani, G., Bagli, E., Chesani, F.: Optilime: Optimized LIME explanations for diagnostic computer algorithms. arXiv preprint arXiv:2006.05714 (2020)
122. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017)
123. Wang, F., Rudin, C.: Falling rule lists. In: Artificial Intelligence and Statistics. pp. 1013–1022 (2015)
124. Watson, D.S., Wright, M.N.: Testing conditional independence in supervised learning algorithms. arXiv preprint arXiv:1901.09917 (2019)

125. Wei, P., Lu, Z., Song, J.: Variable importance analysis: a comprehensive review. Reliability Engineering & System Safety **142**, 399–432 (2015)
126. Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., Wilson, J.: The what-if tool: Interactive probing of machine learning models. IEEE transactions on visualization and computer graphics **26**(1), 56–65 (2019)
127. Williamson, B.D., Feng, J.: Efficient nonparametric statistical inference on population feature importance using Shapley values. arXiv preprint arXiv:2006.09481 (2020)
128. Zeileis, A., Hothorn, T., Hornik, K.: Model-based recursive partitioning. Journal of Computational and Graphical Statistics **17**(2), 492–514 (2008)
129. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)
130. Zhang, Q., Nian Wu, Y., Zhu, S.C.: Interpretable convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8827–8836 (2018)
131. Zhou, Q., Liao, F., Mou, C., Wang, P.: Measuring interpretability for different types of machine learning models. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 295–308 (2018)
132. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology) **67**(2), 301–320 (2005)