


nature machine intelligence

[View all Nature Research journals](#)[Search](#) [Login](#) [Explore](#) [Journal info](#) [Sign up for alerts](#) [RSS feed](#)[nature](#) > [nature machine intelligence](#) > [articles](#) > [article](#)[Article](#) | [Published: 17 January 2020](#)

From local explanations to global understanding with explainable AI for trees

[Scott M. Lundberg](#), [Gabriel Erion](#), [Hugh Chen](#), [Alex DeGrave](#), [Jordan M. Prutkin](#), [Bala Nair](#), [Ronit Katz](#), [Jonathan Himmelfarb](#), [Nisha Bansal](#) & [Su-In Lee](#) 

Nature Machine Intelligence **2**, 56–67(2020)

4726 Accesses | **67** Citations | **91** Altmetric | [Metrics](#)

Abstract

Tree-based machine learning models such as random forests, decision trees and gradient boosted trees are popular nonlinear predictive models, yet comparatively little attention has been paid to explaining their predictions. Here we improve the interpretability of tree-based models through three main contributions. (1) A polynomial time algorithm to compute optimal explanations based on game theory. (2) A new type of explanation that directly measures local feature interaction effects. (3) A new set of tools for understanding global model structure based on combining many local explanations of each prediction. We apply these tools to three medical machine learning problems and show how combining many high-quality local explanations allows us to represent global structure while retaining local faithfulness to the original model. These tools enable us to (1) identify high-magnitude but low-frequency nonlinear mortality risk factors in the US population, (2) highlight distinct population subgroups with shared risk characteristics, (3)

identify nonlinear interaction effects among risk factors for chronic kidney disease and (4) monitor a machine learning model deployed in a hospital by identifying which features are degrading the model's performance over time. Given the popularity of tree-based machine learning models, these improvements to their interpretability have implications across a broad set of domains.

A [preprint version](#) of the article is available at ArXiv.

 [Access through your institution](#)

[Buy or subscribe](#)

Access options

Rent or Buy article

Get time limited or full article
access on ReadCube.

from **\$8.99**

[Rent or Buy](#)

All prices are NET prices.

Subscribe to Journal

Get full journal access for
1 year

\$99.00

only \$8.25 per issue

[Subscribe](#)

All prices are NET prices.
VAT will be added later in the checkout.

Additional access options:

[Log in](#)

[Access through your institution](#)

[Learn about institutional subscriptions](#)

Data availability

The pre-processed mortality data are available at <http://github.com/suinleelab/treexplainer-study>. Privacy restrictions prevent the release of the hospital procedure-related data, and the kidney disease data are only available directly from the National Institute of Diabetes, Digestive and Kidney Diseases (NIDDK).

Code availability

Code supporting this paper is published online at <https://github.com/suinleelab/treexplainer-study>. A widely used Python implementation of TreeExplainer is available at <https://github.com/slundberg/shap>, and portions of it are included in the standard release of XGBoost (<https://xgboost.ai>), LightGBM (<https://github.com/Microsoft/LightGBM>) and CatBoost (<https://catboost.ai>).

References

1. The state of data science & machine learning. *Kaggle* <https://www.kaggle.com/surveys/2017> (2017).
2. Friedman, J., Hastie, T. & Tibshirani, R. *The Elements of Statistical Learning* Vol. 1 (Springer Series in Statistics, Springer, 2001).
3. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4768–4777 (2017).
4. Saabas, A. treeinterpreter python package. *GitHub* <https://github.com/andosa/treeinterpreter> (2019).
5. Ribeiro, M. T., Singh, S. & Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (ACM, 2016).

- 6.** Datta, A., Sen, S. & Zick, Y. Algorithmic transparency via quantitative input influence: theory and experiments with learning systems. In *Proc. 2016 IEEE Symposium on Security and Privacy (SP)*, 598–617 (IEEE, 2016).
- 7.** Štrumbelj, E. & Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* **41**, 647–665 (2014).
- 8.** Baehrens, D. et al. How to explain individual classification decisions. *J. Mach. Learn. Res.* **11**, 1803–1831 (2010).
- 9.** Shapley, L. S. A value for n -person games. *Contrib. Theor. Games* **2**, 307–317 (1953).
- 10.** Sundararajan, M. & Najmi, A. The many Shapley values for model explanation. Preprint at <https://arxiv.org/abs/1908.08474> (2019).
- 11.** Janzing, D., Minorics, L. & Blöbaum, P. Feature relevance quantification in explainable AI: a causality problem. Preprint at <https://arxiv.org/abs/1910.13413> (2019).
- 12.** Matsui, Y. & Matsui, T. NP-completeness for calculating power indices of weighted majority games. *Theor. Comput. Sci.* **263**, 305–310 (2001).
- 13.** Fujimoto, K., Kojadinovic, I. & Marichal, J.-L. Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. *Games Econ. Behav.* **55**, 72–99 (2006).
- 14.** Ribeiro, M. T., Singh, S. & Guestrin, C. Anchors: high-precision model-agnostic explanations. In *Proc. AAAI Conference on Artificial Intelligence* (2018).
- 15.** Shortliffe, E. H. & Sepúlveda, M. J. Clinical decision support in the era of artificial intelligence. *JAMA* **320**, 2199–2200 (2018).
- 16.**

16. Lundberg, S. M. et al. Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760 (2018).

17. Cox, C. S. et al. Plan and operation of the NHANES I Epidemiologic Followup Study, 1992. *Vital Health Stat.* **35**, 1–231 (1997).

18. Chen, T. & Guestrin, C. Xgboost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016).

19. Haufe, S. et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* **87**, 96–110 (2014).

20. Kim, B. et al. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning* (ICLR, 2018).

21. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T. & Lipson, H. Understanding neural networks through deep visualization. In *ICML Deep Learning Workshop* (ICML, 2015).

22. Bau, D., Zhou, B., Khosla, A., Oliva, A. & Torralba, A. Network dissection: quantifying interpretability of deep visual representations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 6541–6549 (IEEE, 2017).

23. Leino, K., Sen, S., Datta, A., Fredrikson, M. & Li, L. Influence-directed explanations for deep convolutional networks. In *Proc. 2018 IEEE International Test Conference (ITC)* 1–8 (IEEE, 2018).

24. Group, S. R. A randomized trial of intensive versus standard blood-pressure control. *N. Engl. J. Med.* **373**, 2103–2116 (2015).

25.

25. Mozaffarian, D. et al. Heart disease and stroke statistics-2016 update a report from the American Heart Association. *Circulation* **133**, e38–e48 (2016).

26. Bowe, B., Xie, Y., Xian, H., Li, T. & Al-Aly, Z. Association between monocyte count and risk of incident CKD and progression to ESRD. *Clin. J. Am. Soc. Nephrol.* **12**, 603–613 (2017).

27. Fan, F., Jia, J., Li, J., Huo, Y. & Zhang, Y. White blood cell count predicts the odds of kidney function decline in a Chinese community-based population. *BMC Nephrol.* **18**, 190 (2017).

28. Zinkevich, M. Rules of machine learning: best practices for ML engineering (2017).

29. van Rooden, S. M. et al. The identification of Parkinson's disease subtypes using cluster analysis: a systematic review. *Mov. Disord.* **25**, 969–978 (2010).

30. Sørbye, T. et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA* **100**, 8418–8423 (2003).

31. Lapuschkin, S. et al. Unmasking clever hans predictors and assessing what machines really learn. *Nat. Commun.* **10**, 1096 (2019).

32. Pfungst, O. Clever Hans: (the Horse of Mr. Von Osten.) A Contribution to Experimental Animal and Human Psychology (Holt, Rinehart and Winston, 1911).

33. *Machine Learning Recommendations for Policymakers* (IIF, 2019);

<https://www.iif.com/Publications/ID/3574/Machine-Learning-Recommendations-for-Policymakers>

34. Deeks, A. The judicial demand for explainable artificial intelligence. (2019).

35.

- Bumb, G., Molitor, D. & Talwalkar, A. S. Model agnostic supervised local explanations. *Adv. Neural Inf. Process. Syst.* **31**, 2520–2529 (2018).
- 36.** Young, H. P. Monotonic solutions of cooperative games. *Int. J. Game Theor.* **14**, 65–72 (1985).
- 37.** Ancona, M., Ceolini, E., Oztireli, C. & Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. In *Proc. 6th International Conference on Learning Representations (ICLR 2018)* (2018).
- 38.** Hooker, S., Erhan, D., Kindermans, P.-J. & Kim, B. A benchmark for interpretability methods in deep neural networks. In *Conference on Neural Information Processing Systems* (NIPS, 2019).
- 39.** Shrikumar, A., Greenside, P., Shcherbina, A. & Kundaje, A. Not just a black box: learning important features through propagating activation differences. Preprint at <https://arxiv.org/abs/1605.01713> (2016).
- 40.** Lunetta, K. L., Hayward, L. B., Segal, J. & Van Eerdewegh, P. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet.* **5**, 32 (2004).
- 41.** Jiang, R., Tang, W., Wu, X. & Fu, W. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics* **10**, S65 (2009).

Acknowledgements

We are grateful to R. Chen, A. Okeson, C. Robinson, V. Khotilovich, N. Hiranuma, J. Janizek, M. T. Ribeiro, J. Schreiber, P. Hall and members of S.-I.L.'s group for the feedback and assistance they provided during the development and preparation of this research. This work was funded by the National Science Foundation (DBI-1759487, DBI-1552309, DBI-1355899, DGE-1762114 and

DGE-1256082), American Cancer Society (127332-RSG-15-097-01-TBG), National Institutes of Health (R35 GM 128638 and R01 NIA AG 061132), and an unrestricted gift from the Northwest Kidney Centers to the University of Washington Kidney Research Institute. The Chronic Renal Insufficiency Cohort (CRIC) study was conducted by the CRIC investigators and supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The data from the CRIC study reported here were supplied by the NIDDK Central Repositories. This manuscript was not prepared in collaboration with Investigators of the CRIC study and does not necessarily reflect the opinions or views of the CRIC study, the NIDDK Central Repositories or the NIDDK.

Author information

Affiliations

1. Microsoft Research, Redmond, WA, USA

Scott M. Lundberg

2. Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave & Su-In Lee

3. Medical Scientist Training Program, University of Washington, Seattle, WA, USA

Gabriel Erion & Alex DeGrave

4. Division of Cardiology, Department of Medicine, University of Washington, Seattle, WA, USA

Jordan M. Prutkin

5. Department of Anesthesiology and Pain Medicine, University of Washington, Seattle, WA, USA

Bala Nair

6. Harborview Injury Prevention and Research Center, University of Washington, Seattle, WA, USA

Bala Nair

7. Kidney Research Institute, Division of Nephrology, Department of Medicine, University of Washington, Seattle, WA, USA

Ronit Katz, Jonathan Himmelfarb & Nisha Bansal

Contributions

S.M.L. and S.I.L. conceived the study. S.M.L. designed algorithms, designed visualizations, designed metrics, ran experiments and contributed to the writing. G.E. ran experiments, designed visualizations and contributed to the writing. H.C. designed algorithms, ran experiments and contributed to the writing. A.D. performed dataset creation. R.K., J.H. and N.B. did dataset selection, model vetting and defined the chronic kidney disease prediction problem. J.M.P., B.N., R.K., J.H. and N.B. each contributed writing and helped procure and interpret datasets. S.-I.L. supervised research, method development and contributed to the writing.

Corresponding author

Correspondence to [Su-In Lee](#).

Ethics declarations

Competing interests

The authors declare no competing interests.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information

Supplementary Information

Supplementary Figs, methods and references.

Supplementary Data 1

Rights and permissions

[Reprints and Permissions](#)

About this article

Cite this article

Lundberg, S.M., Erion, G., Chen, H. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* **2**, 56–67 (2020). <https://doi.org/10.1038/s42256-019-0138-9>

Received 27 June 2019 **Accepted** 06 December 2019 **Published** 17 January 2020

Issue Date January 2020 **DOI** <https://doi.org/10.1038/s42256-019-0138-9>

Subjects [Computer science](#) • [Medical research](#) • [Software](#)

Further reading

- **[Integrating satellite imagery and environmental data to predict field-level cane and sugar yields in Australia using machine learning](#)**

Yuri Shendryk, Robert Davy & Peter Thorburn

Field Crops Research (2021)

- **[Model exploration using conditional visualization](#)**

Catherine B. Hurley

WIREs Computational Statistics (2021)

- **[Enhancing the Evaluation and Interpretability of Data-Driven Air Quality Models](#)**

Jiajun Gu, Bo Yang[...] & K. Max Zhang

Atmospheric Environment (2021)

- **[Automatic detection and characterization of quantitative phase images of thalassemic red blood cells using a mask region-based convolutional neural network](#)**

Yang-Hsien Lin, Ken Y.-K. Liao & Kung-Bin Sung

Journal of Biomedical Optics (2020)

- **Design of an Accurate Machine Learning Algorithm to Predict the Binding Energies of Several Adsorbates on Multiple Sites of Metal Surfaces**

C. S. Praveen & Aleix Comas-Vives

ChemCatChem (2020)

Nature Machine Intelligence ISSN 2522-5839 (online)

© 2020 Springer Nature Limited