Explainable AI for the natural sciences

Joonas Kukkonen, Adiel Lindroos, Anna Brauer

December 23, 2022

1 Introduction

When validating the performance of machine learning (ML) models, most contemporary research focuses on the accuracy of the models. This approach, while useful in its simplicity, makes it impossible to make any fundamental observations of the model outside of whether the model performs well on a particular set of validation data. Together with the rising popularity of highly complex machine learning models, such as neural networks, we have found ourselves surrounded with accurate and powerful machine learning models, yet lacking in capabilities to understand their decisions and functionality. Focusing on performance, other important considerations such as robustness and trust in the systems have been neglected due to a lack of suitable means.

The field of Explainable Artificial Intelligence (XAI) has lately gained traction to meet the demand of understanding the decisions of machine learning models more deeply. XAI methods provide tools to understand why particular models make certain decisions and offer insights into what the models have learned from the underlying data. Additionally, XAI encompasses methods for developing inherently understandable, yet powerful models. While canonical examples and use cases of XAI have focused on fields with high-stakes decision making, such as banking or healthcare, we argue that it is also the natural sciences that can benefit from these methods.

In this paper, we present an overview of XAI and current key approaches. We introduce examples of XAI methods used in natural sciences and their benefits in the respective use cases. Additionally, the role of XAI in natural sciences and virtual laboratories will be discussed.

2 Overview of explainable AI

As the main purpose of XAI is to provide methods and tools to explain artificial intelligence (AI) or ML models and why or how they make certain kinds of predictions or recommendations, there is a need to understand first what explaining a model means. The terminology related to XAI varies a lot; however, there are a few terms that appear frequently. These are interpretability, explainability, and transparency. The definition of these terms varies a lot depending on the source, and the definitions of these terms often mix with each other. However, we still attempt to summarize the terms in the following.

Interpretability means presenting something in understandable terms [4]. In the context of AI or ML, interpretability is defined as presenting something in understandable terms to humans. To evaluate what is humanly understandable is difficult because different users experience things differently, and the experience is subjective [7].

Guidotti et al. [6] defined three different dimensions related to interpretability. These are global versus local, time limitations, and the nature of user experience. Global interpretability means that the user is enabled to understand the whole logic behind the model. By understanding this, the user can understand the logic behind all possible outcomes. In the case of local interpretability, the user understands the reasons behind the individual outcomes of the model. Time limitation relates to the user's time limit to understand the model or outcome. For example, if the user needs to make a decision quickly based on the model's prediction, they need to understand the reasons quickly. On the other hand, if users have time, the explanations can be different. User knowledge or background impacts strongly which kind of explanations they understand; i.e., experienced data scientists understand more complex explanations than a common user may understand.

Explainability relates to the interfaces between a human and the model itself [6]. This explanation should be both understandable to a human and an accurate proxy of the model. Transparency relates to models that are by nature understandable [1]. A human can understand a transparent model and its logic without needing additional explanations.

XAI methods can be divided into two major categories: transparent design and post-hoc methods [1, 6]. In transparent design, models are either simple enough for humans to understand or built from scratch to be understandable by design. Post-hoc methods are methods that are applied to already fitted complex models.

3 Post-hoc approaches: Explaining the black box

Post-hoc approaches are used to explain complex models that are not transparent by themselves [20]. Post-hoc approaches can be divided into two categories: local and global. Local methods explain individual predictions, whereas global methods explain the whole model and its logic. Post-hoc methods, both local and global, can either be model-specific or model-agnostic. Model-agnostic methods can be used with any model to extract information about its inner workings.

3.1 Local approaches

Model-agnostic techniques can be divided into three different approaches: simplification, feature relevance, and visualisation [1]. In simplification methods, a new simplified model is trained to explain a trained black box model. The core of this more simplified model is to reduce the complexity of the model and at the same time retain the performance of the model. Most of these methods are based on rule extraction techniques. A well-known technique of this approach is Local Interpretable Model-Agnostic Explanation (LIME), which we will present later in more detail. Feature relevance methods are based on extracting, measuring, ranking, and presenting the importance of each feature for each prediction. Visual explanation techniques are tools to visually explain black box models.

LIME is a technique for explaining predictions of classifiers and regression models [19]. Its purpose is to explain how a complex model makes a prediction for a specific instance. This is done by training a surrogate model by using subsets of the original input dataset. This interpretable surrogate model should be a good approximation of local predictions, but not the global model itself.

Mashiro et al. [20] used LIME in their study to explain an ML model used as a species distribution model (SDM). Species distribution models are used to predict species distribution variations in space and time based on the relationship between environmental variables and species occurrence data. They used LIME to explain why their SDM model of African elephants made such predictions. The authors used a random forecast algorithm for predicting the distribution and LIME was used to evaluate the reliability of the model. Additionally, they were able to analyse the importance of variables at the local level.

3.2 Global approaches

Global methods aim to answer the question of how the trained model makes predictions and explains its functionality. This could also be understood as extracting what the model has learned about the underlying data. Here the scope is more focused on average prediction, and how the different elements of the model affect this. In fact, some global explanation methods can be thought of and constructed as the average of all local explanations using the same model. Common global explainability methods explore model features, gradients, or weights to explain how the model makes predictions based on the data.

Much like local approaches, global methods can also be model-agnostic or model-specific for certain models or types of models, both with benefits and demerits. Neural networks, both widely used and inherently hard to understand, serve as a great example of a type of model where model-specific approaches are useful compared to more general model-agnostic methods, as they allow for taking advantage of the specific structure of neural networks. Saliency maps are a common method in image processing to explain areas of relevance for a prediction, but this approach can also be extended to global scale and beyond just images. One such example would be a study using *Layer-wise Relevance Propagation* (LRP) [15, 21] to extract climate signals for global temperature maps to better understand what region of the earth is most relevant for predicting decades [11]. LRP functions by back-propagating the relevance of a prediction back to the input layer, producing a result of size of the inputs with relevance values for all input elements. This can be considered as a map of how relevant every input element was to the final output, in this case, the element being a coordinate region, but this could also be pixel if applied to an image, or vector element. Applied to a single prediction, the method is local, but averaged over the data set it produces a global map of the most relevant regions. Indirectly, this method gives us clues on which regions might be most affected by climate change or otherwise relevant to major climate signals. This kind of method allows for examining what regions or variables the model considers most relevant, allowing for rudimentary validation, such as whether the observed trends follow the existing understanding of the phenomenon in question.

In terms of model-agnostic approaches, SHAP [13] is a fairly commonly used method. This is a game theory-centric approach based on calculating Shapley values for predictions. These model the average marginal contribution of a feature over all possible coalitions. This produces a measure of how much a feature contributed to a prediction, compared to the average prediction. Shapley values are also additive, meaning predictions can be combined into a global model. Xue et al. demonstrate the benefits of SHAP and XAI in their paper about modeling Great Lakes surface temperatures [23]. SHAP allows them to validate their Long Short-Term memory neural network model and use this knowledge to combine it with an existing physics-based simulation, mitigating the downsides in both of the models.

4 A priori approaches: Designing interpretable models

Transparent models, i.e., models that are interpretable by design, can be compared to boxes made of glass: the inference process of these models is designed to be clear and transparent for the user. As opposed to post-hoc approaches in which designing the model and explaining it are two different tasks, model design and interpretation are melted into one single step in transparent model design. In other words, the interpretations are intrinsic to the transparent model; the model is a priori interpretable [18].

Different approaches toward transparent model design have been presented in the literature, including data-agnostic approaches based on model simplification or augmentation [6], and domain knowledge-driven model design [18].

4.1 Data-agnostic transparent design

Structurally simple models, e.g., sets of rules and decision trees, are models that humans can understand intuitively [6]. However, to obtain models that are truly transparent by design, it is important to limit the models' complexity; even a decision tree can be inconceivable if it has thousands of branches. For example, Lakkaraju et al. [12] proposed a mechanism for creating interpretable sets of if-then rules by simplifying rules obtained through association rule mining. The mechanism optimizes the rules to be as concise as possible; further objectives include minimizing the overlap between rules and maximizing their coverage and thus relevancy. Interpretable sets of rules are a traditional approach to explainable decision models with applications in, e.g., medical diagnosis [16]. More recently, interpretable rules were used, for example, to gain insights about the interrelations between the human microbiome and the development of diseases [2].

For some applications, inherently transparent types of models are not suitable. In these cases, another option for increasing the model's interpretability is augmentation, i.e., enhancing the model with additional information. A prime example for this strategy is prototype-based reasoning, which is a way to explain the output of classification and clustering methods. Each cluster or class is augmented with a prototype, i.e., the element of the cluster or class that is most typical for the whole group. An example for prototype-based reasoning is the Bayesian case model presented by Kim et al. [10], which learns a representative prototype for each cluster of a Gaussian mixture model. In addition to the protoype, each cluster is annotated with a subspace containing the features that are most important for sorting an element into this particular cluster. Prototype-based reasoning produces powerful explanations because it resembles an essential human reasoning strategy: reasoning by example. For example, Carrizosa et al. [3] showed how prototypes can increase the understanding of cities clustered according to their climate. It should be noted that depending on the definition used, prototype-based reasoning can be classified as a post-hoc explanation method; in this paper, we applied the categorization by Guidotti et al. [6].

4.2 Domain knowledge-driven transparent design

Incorporating domain knowledge enables the transparent design of more complex models, particularly neural networks. There is no one-fits-all approach for this; instead, each model has to be conceptualized individually for its concrete use case. However, Quinn et al. [18] outlined a number of central concepts that can guide the design of transparent models. One important concept is feature engineering, as any model can hardly be interpretable without interpretable, meaningful features created during the pre-processing phase. Furthermore, limiting model complexity can help to increase the model's interpretability. In the case of a neural network, for example, complexity can be reduced by restricted the range of values that the network is allowed to learn. Alternatively, constraints can be enforced on the network's connectivity, e.g., by pruning less relevant connections after training the model.

A third strategy for achieving interpretability is modularization, i.e., partitioning the model into small, meaningful parts. Each module corresponds to a certain physical entity, phenomenon, or system, or complies with human reasoning in another way. To maximize interpretability, the modules should be organized in a clearly understandable structure, e.g., sequentially or in hierarchical order.

Building domain knowledge-based transparent neural networks is a fairly novel concept; in the field of biology, Ma et al. [14] were one of the first to propose a biologically informed neural network, modeling the relationship between the growth rate of yeast and its genotype [17]. An even more recent example of a transparent neural network is P-NET, a deep learning model for predicting the cancer state of prostate cells [5]. In the prediction and early discovery of cancer, explainability is not only a critical factor for patient care but it can also support the investigation of the underlying biological processes. P-NET led to the identification of cancer-predicting genes that were not previously associated with the progression of cancer. This was possible due to the molecularly interpretable architecture of the model. In P-NET, each node of the lowest network layer correspond to a gene; nodes in higher layers correspond to increasingly complex, manually curated biological pathways and systems. When a patient's profile is given as input to the model, the nodes that are activated tell the researchers precisely which genes and pathways influenced the model's decision. This biologically-informed, constrained architecture does not have an adverse effect on accuracy, on the contrary: P-NET outperformed a fully-connected deep neural network.

5 Discussion and conclusion

XAI can be a good additional tool for virtual laboratories and the natural sciences, building trust in the models and providing the capabilities to validate the models. XAI could help bridge the gaps between the different user groups of virtual laboratories, from data scientists and ML experts to domain experts and researchers, providing appropriate tools for all user groups to understand more about the process.

XAI provides robust methods for validating existing ML models. Accuracy as a general metric is useful but limited in its scope and unable to answer questions outside of whether the models perform acceptably on the know data. XAI methods, on the other hand, allow for more general judgments of the model. Examining the features and their importance, it is possible to gauge what the model considers the most important information. Contrasting this information with pre-existing knowledge of the subject matter allows users to make educated guesses about whether the model has learned something useful, and more critically, if it has learned something undesirable.

When the model performs as expected, XAI is able to prove its correctness and thus builds trust in the system. The scientific method is traditionally concerned with transparency and trustworthiness; a theory or method should be proven before acceptance. It could be argued that this principle should also be applied to the use of ML models in natural sciences. The functionality of the model and its trustworthiness should be proven more rigorously than in current common practice, for which XAI could provide methods.

The potential of XAI for generating new scientific insights should not be underestimated either, even though this is not the principal purpose of XAI. Integrating XAI into virtual laboratories could open opportunities for exploratory science, giving researchers the possibility to examine model decisions thoroughly and discover unexpected patterns. On their own, XAI techniques cannot prove physical phenomena, but they can support researchers in uncovering aspects worth studying more in-depth.

XAI is a comparably young field, as the concept gained traction only in the last few years. Although some methods for explaining machine learning models have been around much longer, the awareness of the need for integrating explanation methods into the workflow increased only recently [8]. As a consequence, general-purpose XAI tools are not yet widely established; these, however, are needed to facilitate the smooth integration of XAI in virtual laboratories.

Explanations generated with XAI methods are not immune to errors, uncertainties, and misinterpretations. To avoid the latter, it is crucial to find effective ways of communicating explanations to the end users, i.e., explanation interfaces (e.g., 22). Considering the users of a full virtual laboratory, who may have significantly varying professional backgrounds, designing these interfaces will be a challenge on its own. A part of this problem is also that it would technically be necessary to communicate and handle explanation uncertainties. Quantifying the uncertainty of explanations is not a well-researched topic, which further contributes to this challenge.

Comparing the different types of XAI methods, we could argue that solutions that are transparent by design should always be preferred in a virtual laboratory, as the risk of misinterpretations is generally lower. However, as transparent model design is much less generic compared to post-hoc methods, it requires more resources. Additionally, it is not always possible or feasible, and still a very new concept overall. For example, Jiménez-Luna et al. [9] found in 2020 that no transparent solutions have been proposed for the fields of chemistry and drug design yet. At the same time, they recognize how transparent models could help solve many types of problems in drug discovery, stating even that interpretable models could provide the means for replacing animal testing by extrapolating from in vitro to in vivo experiments.

To conclude, it is clear that the potential of incorporating XAI into the scientific process is enormous. Ideally, it should become a fixed part of every research project using ML models, and given the rapid current progress, there is a chance that it will establish itself rather sooner than later. In the virtual laboratories of the future, XAI can be expected to fill different roles, including but not limited to guiding model design, enhancing model validation, providing explanations of model decisions, and serving as a tool for exploratory science.

References

[1] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58:82–115, 2020.

[2] E. Bogart, R. Creswell, and G. K. Gerber. MITRE: inferring features from microbiota time-series data linked to host status. *Genome biology*, 20(1):1–15, 2019.

[3] E. Carrizosa, K. Kurishchenko, A. Marín, and D. R. Morales. Interpreting clusters via prototype optimization. *Omega*, 107:102543, 2022.

[4] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 27, 2017.

[5] H. A. Elmarakeby, J. Hwang, R. Arafeh, J. Crowdis, S. Gang, D. Liu, S. H. Al-Dubayan, K. Salari, S. Kregel, C. Richter, et al. Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598(7880):348–352, 2021.

[6] 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.

[7] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang. XAI—explainable artificial intelligence. *Science robotics*, 4(37):eaay7120, 2019.

[8] R. Hughes, C. Edmond, L. Wells, M. Glencross, L. Zhu, and T. Bednarz. eXplainable AI (xai) an introduction to the XAI landscape with practical examples. In *SIGGRAPH asia 2020 courses*, pages 1–62. 2020.

[9] J. Jiménez-Luna, F. Grisoni, and G. Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584, 2020.

[10] B. Kim, C. Rudin, and J. A. Shah. The bayesian case model: A generative approach for case-based reasoning and prototype classification. *Advances in neural information processing systems*, 27, 2014.

[11] Z. M. Labe and E. A. Barnes. Detecting climate signals using explainable AI with single-forcing large ensembles. *Journal of Advances in Modeling Earth Systems*, 13 (6):e2021MS002464, 2021.

[12] H. Lakkaraju, S. H. Bach, and J. Leskovec. 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, 2016.

[13] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[14] J. Ma, M. K. Yu, S. Fong, K. Ono, E. Sage, B. Demchak, R. Sharan, and T. Ideker. Using deep learning to model the hierarchical structure and function of a cell. *Nature methods*, 15(4):290–298, 2018.

[15] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller. *Layer-Wise Relevance Propagation: An Overview*, pages 193–209. Springer International Publishing, 2019. ISBN 978-3-030-28954-6.

[16] D. Nauck and R. Kruse. Obtaining interpretable fuzzy classification rules from medical data. *Artificial intelligence in medicine*, 16(2):149–169, 1999.

[17] G. Novakovsky, N. Dexter, M. W. Libbrecht, W. W. Wasserman, and S. Mostafavi. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, pages 1–13, 2022.

[18] T. P. Quinn, S. Gupta, S. Venkatesh, and V. Le. A field guide to scientific XAI: Transparent and interpretable deep learning for bioinformatics research. *arXiv preprint arXiv:2110.08253*, 2021.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin. ”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, 2016.

[20] M. Ryo, B. Angelov, S. Mammola, J. M. Kass, B. M. Benito, and F. Hartig. Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography*, 44(2):199–205, 2021.

[21] B. Toms, E. Barnes, and I. Ebert-Uphoff. Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, 12, 09 2020.

[22] C.-H. Tsai and P. Brusilovsky. Designing explanation interfaces for transparency and beyond. In *IUI Workshops*, 2019.

[23] P. Xue, A. Wagh, G. Ma, Y. Wang, Y. Yang, T. Liu, and C. Huang. Integrating deep learning and hydrodynamic modeling to improve the great lakes forecast. *Remote Sensing*, 14(11), 2022.