Multiple stakeholders drive diverse interpretability requirements for machine learning in healthcare

Fergus Imrie[1*], Robert Davis[2] and Mihaela van der Schaar[2,3]

[1]University of California, Los Angeles, USA.
[2]University of Cambridge, UK.
[3]The Alan Turing Institute, UK.

*Corresponding author(s). E-mail(s): imrie@ucla.edu;
Contributing authors: rd473@cam.ac.uk; mv472@cam.ac.uk;

Abstract

Applications of machine learning are becoming increasingly common in medicine and healthcare, enabling more accurate predictive models. However, this often comes at the cost of interpretability, limiting the clinical impact of machine learning methods. To realize the potential of machine learning in healthcare, it is critical to understand such models from the perspective of multiple stakeholders and various angles, necessitating different types of explanations. In this perspective, we motivate and explore five fundamentally different types of *post hoc* machine learning interpretability. We highlight the different information they provide and describe when each can be useful. We examine the various stakeholders in healthcare, delving into their specific objectives, requirements, and goals. We discuss how current notions of interpretability can help meet these and what is required for each stakeholder to make machine learning models clinically impactful. Finally, to facilitate adoption, we release an open-source interpretability library (https://github.com/vanderschaarlab/Interpretability) containing implementations of the different types of interpretability, including tools for visualizing and exploring the explanations.

Keywords: Machine learning, Interpretability, Explainable AI, Healthcare, Medicine

Introduction

Machine learning approaches are being increasingly proposed for predictive modeling in medicine and healthcare and have the potential to revolutionize medicine and become core clinical tools [1]. However, machine learning methods have failed to make significant translational impact thus far, with very few artificial intelligence (AI) systems currently in clinical deployment. Furthermore, without due care, AI approaches have the potential to be overused or misused, possibly causing patient harm [2]. For advances in machine learning and AI to be clinically actionable and capable of making real-world impact, the methods proposed by the machine learning community must be more than highly predictive. Instead, users must be able to understand and debug how models issue predictions, and models should provide insight to further medical knowledge.

Without a transparent understanding of how models make predictions, they may act in unintended and undesirable ways. For example, models may learn incorrect or aberrant features unique to

the training data, leading to biased or unfair decisions [3, 4]. Indeed, uninterpretable models pose a threat to medical ethics, with possible detrimental consequences for both individual and public health [5]. Understanding and debugging machine learning systems are critical steps to build model trust [6] and necessary for medical professionals and the broader public before clinical deployment [7]. Furthermore, a clear understanding of computational models is now a regulatory requirement for deployment in many healthcare systems globally [8, 9].

An additional challenge in medicine is the presence of multiple stakeholder groups, namely model developers, medical researchers, legislators and regulators, clinicians, and patients, each with their own backgrounds, expertise, and goals. As a result, individuals from each group may need different types of explanations. For example, a clinician may want to know which features are typically most important for the AI system when issuing predictions, allowing them to understand better the underlying logic behind the system. On the other hand, a patient may primarily be interested in understanding why an AI system issued the prediction for them and what changes they might be able to make to change their prognosis.

Beyond inherently interpretable models

One approach to ensure machine learning models are not only predictive but can also be understood is to adopt inherently interpretable models [10]. Typically, such models are either explicit in their functional form (e.g., linear regression) or about the logical rules used to issue a prediction (e.g., decision tree). However, there are limitations to this approach. First, black-box machine learning approaches have become de facto approaches in several domains such as image analysis [11, 12], natural language processing [13, 14], and multimodal learning [15], and have been shown to exhibit improved performance compared to white-box models in some other circumstances [16–18], although this is far from always the case [19, 20]. Second, the rationale for predictions could still be complex or not readily understood, even for such models, leading to the interpretability requirements of all stakeholders not being met. Thus, techniques that allow us to study models in a *post hoc* manner are particularly important and desirable. It has been argued that some opaqueness is acceptable and, further, that improved predictive accuracy is more important than being able to explain how a system achieved it [21]. While there is merit in this argument, we believe that we should strive for a deeper understanding of machine learning models. Additionally, widespread clinical adoption and acceptance will only be achieved with thorough auditing and comprehension of machine learning models.

In this paper, we motivate and describe the goals and requirements of five key stakeholders in the development and administration of clinical machine learning models. We then explore how different types of explanations can begin to address these diverse needs. We present five distinct types of *post hoc* interpretability, highlight the different information they provide, and describe when each can be useful. As a tool for both the clinical and machine learning communities, we provide an open-source software package containing a suite of interpretability methods and a visualization platform to make a range of explanations readily accessible.

Types of Explanation

Before discussing different types of explanations, we should first define what we mean by interpretable. While we can explicitly write the computation performed by a neural network, this does not make them interpretable. Instead, we follow the definition of Biran and Cotton [22], also adopted by Miller [23] among others, where interpretability is defined as the degree to which the cause of a prediction can be understood. Further, we define explainability as post hoc interpretability [24] and, as a result, we will often use the two terms interchangeably.

Broadly speaking, we can group current *post hoc* explainable AI (XAI) methods into five fundamentally different classes, each of which offers a different type of explanation and has a unique role in understanding and debugging clinical machine learning models. In particular, methods can provide explanations that are feature-based, example-based, concept-based, model-based, or counterfactual (Table 1). We begin with a brief introduction of each.

Feature-based explanations

Feature-based explainability methods are the most common type of XAI and allow the user to understand the importance of each feature. Some feature-based methods explain individual predictions (local), while others provide the global relevance of features. Popular methods for determining feature importance include local methods such as LIME [25], SHAP [26], and Integrated Gradients [27], as well as global methods such as partial dependence plots [28] and permutation feature importance [29, 30]. Understanding the relevance of each feature is helpful in many contexts; as such, feature-based explanations can provide valuable insight throughout the model development cycle. One limitation of such approaches is that they typically try to isolate the importance or impact of a specific feature. However, features often act jointly in combination [31], and this information is not directly provided by feature-importance methods.

Example-based explanations

A second and fundamentally different type of XAI is example-based explanations. Instead of quantifying the relevance of each feature for the model, example-based methods explain predictions by providing the user with other instances, often from the training set, that the model views as being most similar to a given sample. This approach shares similarities with the more general technique of Case-Based Reasoning [38]. Example-based methods include Simplex [32] and ExMatchina [39]. One challenge with example-based methods is understanding *why* the model views two samples as similar. A recent example-based method has tried to address this by bridging feature-based and example-based approaches to explain the importance of each feature in the similar examples [32].

One compelling feature of example-based methods is that they allow the user to customize the explanation to their expertise. This can be particularly important when transferring models to new settings beyond the environment in which the model was developed [40]. For example, by changing which examples are available, a clinician can understand the predictions of a model in terms of patients they know or canonical cases. In addition, example-based methods can be used to debug models. First, if the explanation reveals that the model views two samples as similar, but the user disagrees, this could indicate a flaw in the model's logic. Second, if the model incorrectly classifies the most similar examples, this casts doubt on the validity of the prediction.

Concept-based explanations

Concept-based interpretability methods, such as concept activation vectors (CAV) [33] and concept activation regions (CAR) [34], allow users to investigate predictive models and test whether they utilize specific concepts. Given a concept specified by the user via a set of examples (e.g., stripes in an image), a classifier is used to assess whether the internal representations of a model differ between examples where a concept is present and examples where a concept is absent.

In some domains, such as imaging, understanding the importance of individual features (i.e., pixels) might not be insightful since features do not necessarily carry significant meaning alone. In contrast, since the user provided the concept, concept-based explanations are customizable and understandable by design. In addition, concepts can be specified at any level of abstraction and allow the user to probe the model in a fundamentally different way since concepts are not input features to the model. While a strength, that the user must currently provide concepts is also a limitation, although we note the development of approaches for concept discovery [41]. Additionally, in general, concept-based explanations do not take into account how much each concept plays a role in the prediction.

Model-based explanations

While feature-based interpretability methods can explain the predictions of a model with the contributions of individual features, they do not provide deeper insight, such as whether the model is non-linear or interactions between features exist. Model-based explanations use auxiliary models, also known as meta-models, to convert a black-box model into a different form that can be used to analyze what the model has learned.

One model-based method to explain predictions is associative classifiers [42, 43], which learn a set of clinically-interpretable association rules (or "if-then" clauses), similar to a decision tree. An advantage of model-based approaches such as

Table 1 Definition and examples of each of the different classes of explanation methods.

Explanation class	Definition	Examples
Feature-based	Provides the importance of each feature to model predictions	LIME [25], SHAP [26]
Example-based	Explains model predictions with reference to other examples	SimplEx [32]
Concept-based	Explains model predictions with reference to human-defined concepts	CAV [33], CAR [34]
Model-based	Explains model predictions via auxiliary meta-models	Symbolic pursuit [35, 36]
Counterfactual	Explains model predictions by generating synthetic example(s) that are similar but with a different prediction	MOC [37]

associative classifiers is that they not only provide interpretations of the model behavior but also distill clinical insights from the base model's underlying prediction rules (e.g., [44]).

An alternate approach is using symbolic regression to convert a black-box model into a closed-form equation [35, 36]. This approach allows us to elucidate the precise functional forms by which a model captures non-linearities in the data, identifying which features interact and how strongly. In addition, by converting to an equation, we can quantify the importance of each feature, demonstrating a link between the different types of interpretability. Finally, after converting a machine learning model into an equation, we have access to the entire range of mathematical techniques classically used to analyze equations, allowing deep and rigorous analysis not possible for the machine learning model directly. Like other explainability approaches, model-based methods suffer from only being a proxy for the underlying model. In particular for model-based approaches, there is a natural complexity-accuracy tradeoff between the two models.

Counterfactual explanations

Counterfactuals are local explanations that offer insight into the reasoning behind specific predictions by identifying alterations to the input feature values that result in a different model output. Unlike example-based explanations, counterfactuals are typically not actual samples. However, generally, it is desirable for counterfactuals to be plausible samples from the underlying data distribution and to modify the fewest number of features.

Compared to many other explanations, a benefit of counterfactuals is the lack of additional assumptions. In particular, to construct counterfactuals, the user only needs to be able to query the system and receive the output. However, generating counterfactuals is a multi-objective optimization problem and depends on the user's preferences. In addition, counterfactuals are not unique. Generating multiple counterfactual explanations often provides additional insight into the model behavior by uncovering multiple viable ways of altering the original sample to achieve a different prediction. However, this lack of uniqueness can also be a downside since a feature being unaffected does not mean that changing its value would not alter the prediction.

Finally, by interpreting the features as causing a model's prediction, counterfactual explanations can be seen as causal for the model, even if they are not necessarily counterfactuals for the underlying joint distribution. As a result, the interpretation of counterfactuals is clear, and they can be particularly useful in scenarios where it is possible to modify the underlying features via interventions.

Stakeholders

There are multiple stakeholders in medicine, each with different goals and requirements for machine learning systems and explanations. To meet these diverse requirements, we must engage with multiple types of explanations. However, while the

Table 2 Representative questions for each of the key stakeholders in healthcare and which type of explanation can be used to provide the appropriate insight.

Stakeholder	Representative questions	Explanation type
Model Developer	Which features should be included in the model?	Feature-based
	Do the features used by the model make sense?	Feature-based
	Does the model use established medical concepts?	Concept-based
Medical Researcher	Which features are predictive?	Feature-based
	What are the relationships between features?	Model-based
	Does this feature provide additional information?	Feature-based
	What is the impact of changing this feature?	Counterfactual
Regulator	What features does the model use to issue its predictions?	Feature-based
	Does the model understand known medical concepts?	Concept-based
	What explicit equations could be used instead?	Model-based
	Does the model see these patients as similar?	Example-based
	What is the impact of changing this feature? Does it agree with current understanding?	Counterfactual
Clinician	What features does the model use to issue its predictions?	Feature-based
	Does the model understand known medical concepts?	Concept-based
	What previous patients is this patient similar to?	Example-based
	What effect would changing this characteristic have on the prediction?	Counterfactual
Patient	What characteristics led to the predictions?	Feature-based
	What other patients are similar?	Example-based
	What can I change to alter the prediction?	Counterfactual

machine learning community has developed a diverse range of explanations, the uptake by the medical community has largely been limited to feature-based explanations.

In this section, we discuss five key stakeholders in healthcare and present their unique perspectives. We then outline how different types of explanations can help address their various goals. Note that the aims of each stakeholder are not mutually exclusive, and individuals may, and often will, adopt multiple roles (e.g., model developer and medical researcher). In addition, the characterization presented herein may not fully capture all views of each stakeholder and instead is intended to be representative of the breadth and nature of interpretability requirements. We provide a set of representative questions for each of the stakeholders together with which type of explainability could be used to address the questions in Table 2.

We hope that explicitly defining a set of key stakeholders and presenting a set of tools facilitates better conversations between the machine learning and clinical communities.

Model developers

We refer to anyone developing a new machine learning model as a "model developer". The primary goal for model developers is to produce highly predictive models. While understanding how a model issues its predictions might not seem strictly necessary for this goal, selecting a subset of features or performing feature engineering using feature-based explainability methods has been shown to improve performance, simplify models, and reduce deployment costs [45, 46].

Separately, model developers need to ensure that models are operating as expected and have the potential to be deployed in the real world. In particular, shortcut learning [47], where the model

learns a spurious relationship, must be avoided, as seen in several machine learning applications in healthcare [4, 48, 49]. This can involve the use of multiple types of XAI. For example, feature-based explanations can be used to check that the most important features are consistent with current scientific knowledge, while concept-based explanations allow developers to check whether the model is using established medical concepts.

Medical researchers

Machine learning models are beginning to be used as a tool for medical and scientific discovery, making machine learning useful even if the models themselves are never deployed. This is achieved by training a machine learning model and then analyzing it using techniques from XAI. Medical researchers can use feature-based explanations to test hypotheses about the importance of features or discover new predictive features. Sometimes, researchers may desire more precise relationships between features. In this case, model-based explanations can elucidate highly-performant black-box models and uncover feature interactions and equations for their relationships. Lastly, counterfactual explanations offer an alternate way to understand the importance and impact of certain features, while enabling researchers to test and form hypotheses about potential interventions.

Medical researchers' use of machine learning, and consequently XAI, may differ from other stakeholders. For example, while optimizing performance may be desired by clinicians and patients, medical researchers instead may seek models that can be used to generate hypotheses or give insight into underlying biological mechanisms.

As a specific example of how XAI is being used to facilitate medical research, take liquid biopsy for cancer screening. While many potential biomarkers can be screened during preliminary sequencing studies, a much smaller number must be used during device implementation due to practical hardware limitations [50]. As a result, XAI has been used to select the most important features and reveal the interactions between selected genes [51].

Regulators

Regulatory checks on machine learning models before deployment in high-stakes medical environments are rightly becoming more onerous and rigorous [52]. For example, a detailed understanding of how computational models function is now a requirement for deployment in many healthcare systems. In the United States, the Food and Drug Administration (FDA) demands "transparency about the function and modifications of medical devices" as a critical safety aspect [8], while in the European Union, Article 22 of GDPR legislation requires that "meaningful information about the logic involved" be provided [9].

Since each type of explanation offers a unique angle, all can be valuable to regulators as they probe and debug machine learning models in healthcare. In particular, regulators can use feature-based methods to understand what variables the model is using to issue predictions, while concept-based approaches can check whether the model respects established medical concepts. For example, fairness and bias are two important considerations in healthcare [53], and existing biases in the data should not be reinforced by models [54]. XAI tools can be used to assess fairness and bias, as well as understand their origin. XAI can be used to understand if these biases occur and to what extent. Beyond this, some regulators may place restrictions on the type of models deemed acceptable. For example, the American Joint Committee on Cancer requires explicit risk equations [55]. Model-based explainability can enable complex machine learning models to be used to discover better explicit equations (e.g. [56]). Finally, regulators can use example-based and counterfactual explanations to probe the model on an instance-wise basis and check for local pathologies in the model.

Clinicians

For clinicians to use complex machine learning models to assist in making high-stakes decisions, they need to trust the predictions being issued. This requires clinicians to understand how the model operates both globally and locally. Global trust can be built via feature-based, concept-based, and model-based explanations, allowing clinicians to gain insight into the general function of the model beyond estimates of performance.

However, a global understanding of a model is insufficient – clinicians need to understand the rationale behind the prediction for the current patient. Instance-wise, or local, feature importance can help clinicians understand what features are driving the prediction for a specific patient, while example-based explanations can relate the current prediction to previous patients or canonical cases. Example-based explanations can be particularly important when debugging models or deciding whether deviate from their predictions. For example, if the examples have incorrect predictions, the clinician might not trust the issued prediction. Alternatively, if the clinician disagrees that the examples are similar, this could indicate an issue with the predictive model. Finally, counterfactuals can help guide clinicians as to what change in the patient's covariates would lead to an improvement or deterioration in their condition or outlook.

On the other hand, distilling more complex models into alternate, simpler forms can aid clinical practice both from a usability and interpretability perspective. For example, an associative classifer was used to extract clinically interpretable rules using 3 variables from a machine learning model utilizing 115 variables [44]. While there is an undoubted tradeoff, the ability to convert arbitrarily complex models into a form that is clinically actionable is of substantial benefit.

One prominent example of XAI is in medical image analysis, where both AI systems and explanations of predictions are becoming increasingly commonplace [57]. When using ML for triage, clinicians require information as to why a patient has been referred, while when AI is used as a second reader, XAI is particularly critical to resolve discrepancies between radiologists and ML predictions. Furthermore, in many cases, for example pneumonia, a clinical diagnosis alone is of limited use and it is important to understand how the imaging should be used in clinical management, such as when and with what types of antibiotics to treat.

Finally, the use of XAI is not limited to explanations to help clinicians understand models, but can be used to facilitate human-machine partnerships and improve clinical outcomes. For example, a study on the diagnosis of tuberculosis on chest X-rays showed that 10 out of the 13 participating physicians had better diagnostic accuracy when assessing chest X-rays with XAI than without [58].

Patients

Patients primarily wish to understand why a prediction was issued for them. Feature-based explanations can identify the patient's characteristics that led to the prediction, while example-based methods can allow the patient to see similar patients and their outcomes. Finally, counterfactual explanations can allow the patient to understand what would need to be different to change the prediction, which could be helpful in circumstances where the patient can influence certain covariates (e.g., smoking, diet, or alcohol consumption). To a lesser extent, patients also wish to understand and trust models more generally. Similar methods to clinicians (see above) can be used.

Other stakeholders

While we discuss five primary medical stakeholders above, this is by no means exhaustive and other stakeholders exist. For example, we have not explored the role of funders of ML systems or industry. While distinct, there is significant overlap between these stakeholders and those we have discussed. Industry often conducts medical research, and before decisions are taken to fund projects such as clinical trials of ML systems, decision makers will want to gain a deep understanding of the model.

Discussion

Significant work is still needed to improve the quality of both machine learning models and explanations before such systems can become commonplace in clinical practice [5, 59]. Like the models they explain, interpretability methods are imperfect; they do not exactly capture precisely how a model works [10, 24] and different approaches of analysing the same model have been shown to lead to different conclusions [60]. Indeed, there have been several criticisms of explainable AI in healthcare, with some arguing that using interpretability methods to understand individual predictions offers "false hope", in part due to an "interpretability gap" [61] resulting from both the

choice of XAI method and how the explanation is interpreted. As a result, Ghassemi et al. cautioned "against having explainability be a requirement for clinically deployed models" [61]. While we agree with this word of warning and acknowledge the issues that have been demonstrated with some explainability methods (e.g., saliency maps [10, 61]), we believe XAI has much to offer the medical community, and others agree [5, 62, 63]. Even Ghassemi et al. concede that "explanations can be extremely useful when applied to global AI processes" [61] and the much-maligned saliency maps have also been successfully used to further medical knowledge [64, 65].

Consequently, techniques that explain machine learning models will prove invaluable for the medical community. So far, applications of interpretability in healthcare have predominantly focused on feature-based approaches. However, to harness the full potential of explainability, other techniques and *types* of interpretability, such as those discussed in this Perspective, are required. In particular, there has not been sufficient engagement and awareness about the capabilities of current tools to explain machine learning models, nor how the multiple different stakeholders in healthcare should use them. While Amann et al. [5] provide a thoughtful discussion on the relevance of explainability from multiple perspectives, they do not offer concrete suggestions and guidance on the capabilities and different types of interpretations that are available. In addition, the role of different stakeholders in healthcare has been overlooked, with studies too often focused on only one aspect. To raise awareness and promote broader consideration, we have described the various stakeholders, highlighting their different roles and requirements. To foster discussion and show the different ways stakeholders can engage with explainable AI, we have detailed the toolbox of methods available for understanding and debugging machine learning models. Finally, we have connected the two, detailing how different stakeholder requirements can be addressed using the XAI toolbox.

We hope this will promote further engagement and collaboration between the machine learning and medical communities. Based on our discussions with healthcare professionals [66], a key impediment to adoption is a lack of a readily-available, easy-to-use implementation of a range of methods to explain machine learning models. Consequently, we have developed an open-source software package containing a suite of interpretability methods and a visualization platform, making a range of explanations accessible to both the medical and machine learning communities. Our open-source software package for machine learning interpretability is provided at https://github.com/vanderschaarlab/Interpretability. For machine learning to realize its potential in healthcare, both communities need to experiment with the available tools and identify missing pieces in the puzzle.

Finally, while we have highlighted the medical setting, consideration of multiple stakeholders is essential in other fields, such as criminal justice, education, and finance.

Acknowledgments. F.I. and M.vdS. are supported by the National Science Foundation (NSF), grant number 1722516. In addition, M.vdS. is supported by the Office of Naval Research (ONR).

Competing interests. The authors declare no competing interests

References

[1] Topol, E.J.: High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine **25**(1), 44–56 (2019). https://doi.org/10.1038/s41591-018-0300-7

[2] Volovici, V., Syn, N.L., Ercole, A., Zhao, J.J., Liu, N.: Steps to avoid overuse and misuse of machine learning in clinical research. Nature Medicine **28**(10), 1996–1999 (2022). https://doi.org/10.1038/s41591-022-01961-6

[3] 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, 1721–1730 (2015)

[4] Winkler, J.K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., Haenssle, H.A.: Association between surgical skin markings in dermoscopic images and

diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatology **155**(10), 1135–1141 (2019). https://doi.org/10.1001/jamadermatol.2019.1735

[5] Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I., the Precise4Q consortium: Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making **20**(1), 310 (2020). https://doi.org/10.1186/s12911-020-01332-6

[6] Rajpurkar, P., Chen, O. Emmaand Banerjee, Topol, E.J.: AI in health and medicine. Nature Medicine **28**(1), 31–38 (2022). https://doi.org/10.1038/s41591-021-01614-0

[7] Yoon, C.H., Torrance, R., Scheinerman, N.: Machine learning in medicine: Should the pursuit of enhanced interpretability be abandoned? Journal of Medical Ethics **48**(9), 581–585 (2022). https://doi.org/10.1136/medethics-2020-107102

[8] Food and Drug Administration and others: Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) (2019)

[9] Mourby, M., Ó Cathaoir, K., Collin, C.B.: Transparency of machine-learning in healthcare: The GDPR & European health law. Computer Law & Security Review **43**, 105611 (2021)

[10] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x

[11] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017). https://doi.org/10.1038/nature21056

[12] Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P.C., Mega, J.L., Webster, D.R.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA **316**(22), 2402–2410 (2016). https://doi.org/10.1001/jama.2016.17216

[13] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. Advances in Neural Information Processing Systems **33**, 1877–1901 (2020)

[14] Jiang, L.Y., Liu, X.C., Nejatian, N.P., Nasir-Moin, M., Wang, D., Abidin, A., Eaton, K., Riina, H.A., Laufer, I., Punjabi, P., Miceli, M., Kim, N.C., Orillac, C., Schnurman, Z., Livia, C., Weiss, H., Kurland, D., Neifert, S., Dastagirzada, Y., Kondziolka, D., Cheung, A.T.M., Yang, G., Cao, M., Flores, M., Costa, A.B., Aphinyanaphongs, Y., Cho, K., Oermann, E.K.: Health system-scale language models are all-purpose prediction engines. Nature (2023). https://doi.org/10.1038/s41586-023-06160-y

[15] Soenksen, L.R., Ma, Y., Zeng, C., Boussioux, L., Villalobos Carballo, K., Na, L., Wiberg, H.M., Li, M.L., Fuentes, I., Bertsimas, D.: Integrated multimodal artificial intelligence framework for healthcare applications. npj Digital Medicine **5**(1), 149 (2022). https://doi.org/10.1038/s41746-022-00689-4

[16] Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G.E., Irvine, J., Le, Q., Litsch, K., Mossin, A., Tansuwan, J., Wang, D., Wexler, J., Wilson, J., Ludwig, D., Volchenboum, S.L., Chou,

K., Pearson, M., Madabushi, S., Shah, N.H., Butte, A.J., Howell, M.D., Cui, C., Corrado, G.S., Dean, J.: Scalable and accurate deep learning with electronic health records. npj Digital Medicine **1**(1), 18 (2018). https://doi.org/10.1038/s41746-018-0029-1

[17] Alaa, A.M., Bolton, T., Di Angelantonio, E., Rudd, J.H.F., van der Schaar, M.: Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLOS ONE **14**(5), 1–17 (2019). https://doi.org/10.1371/journal.pone.0213653

[18] Lee, C., Light, A., Saveliev, E.S., van der Schaar, M., Gnanapragasam, V.J.: Developing machine learning algorithms for dynamic estimation of progression during active surveillance for prostate cancer. npj Digital Medicine **5**(1), 110 (2022). https://doi.org/10.1038/s41746-022-00659-w

[19] Akbilgic, O., Davis, R.L.: The promise of machine learning: When will it be delivered? Journal of Cardiac Failure **25**(6), 484–485 (2019). https://doi.org/10.1016/j.cardfail.2019.04.006

[20] Schulz, M.-A., Yeo, B.T.T., Vogelstein, J.T., Mourao-Miranada, J., Kather, J.N., Kording, K., Richards, B., Bzdok, D.: Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. Nature Communications **11**(1), 4238 (2020). https://doi.org/10.1038/s41467-020-18037-z

[21] London, A.J.: Artificial intelligence and black-box medical decisions: Accuracy versus explainability. Hastings Center Report **49**(1), 15–21 (2019). https://doi.org/10.1002/hast.973

[22] Biran, O., Cotton, C.: Explanation and justification in machine learning: A survey. In: IJCAI-17 Workshop on Explainable AI (XAI), vol. 8, pp. 8–13 (2017)

[23] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)

[24] Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018)

[25] 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)

[26] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems **30** (2017)

[27] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning, pp. 3319–3328 (2017). PMLR

[28] Friedman, J.H.: Greedy function approximation: A gradient boosting machine. The Annals of Statistics **29**(5), 1189–1232 (2001). https://doi.org/10.1214/aos/1013203451

[29] Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[30] 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)

[31] Imrie, F., Norcliffe, A.L.I., Lio, P., van der Schaar, M.: Composite feature selection using deep ensembles. Advances in Neural Information Processing Systems **35**, 36142–36160 (2022)

[32] Crabbe, J., Qian, Z., Imrie, F., van der Schaar, M.: Explaining latent representations with a corpus of examples. Advances in Neural Information Processing Systems **34**, 12154–12166 (2021)

[33] Kim, B., Wattenberg, M., Gilmer, J., Cai,

C., Wexler, J., Viegas, F., Sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: International Conference on Machine Learning, pp. 2668–2677 (2018). PMLR

[34] Crabbé, J., van der Schaar, M.: Concept activation regions: A generalized framework for concept-based explanations. Advances in Neural Information Processing Systems **35** (2022)

[35] Alaa, A.M., van der Schaar, M.: Demystifying black-box models with symbolic metamodels. Advances in Neural Information Processing Systems **32** (2019)

[36] Crabbe, J., Zhang, Y., Zame, W.R., van der Schaar, M.: Learning outside the black-box: The pursuit of interpretable models. Advances in Neural Information Processing Systems **33**, 17838–17849 (2020)

[37] Dandl, S., Molnar, C., Binder, M., Bischl, B.: Multi-objective counterfactual explanations. In: International Conference on Parallel Problem Solving from Nature, pp. 448–469 (2020). Springer

[38] Aamodt, A., Plaza, E.: Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Communications **7**(1), 39–59 (1994)

[39] Jeyakumar, J.V., Noor, J., Cheng, Y.-H., Garcia, L., Srivastava, M.: How can I explain this to you? An empirical study of deep neural network explanation methods. Advances in Neural Information Processing Systems **33**, 4211–4222 (2020)

[40] Wiesenfeld, B.M., Aphinyanaphongs, Y., Nov, O.: AI model transferability in healthcare: a sociotechnical perspective. Nature Machine Intelligence **4**(10), 807–809 (2022). https://doi.org/10.1038/s42256-022-00544-x

[41] Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. Advances in Neural Information Processing Systems **32** (2019)

[42] Thabtah, F.: A review of associative classification mining. The Knowledge Engineering Review **22**(1), 37–65 (2007)

[43] Luo, G.: Automatically explaining machine learning prediction results: A demonstration on type 2 diabetes risk prediction. Health information science and systems **4**(1), 1–9 (2016)

[44] Alaa, A.M., van der Schaar, M.: Prognostication and risk factors for cystic fibrosis via automated machine learning. Scientific Reports **8**(1), 11242 (2018). https://doi.org/10.1038/s41598-018-29523-2

[45] Min, F., Hu, Q., Zhu, W.: Feature selection with test cost constraint. International Journal of Approximate Reasoning **55**(1, Part 2), 167–179 (2014). https://doi.org/10.1016/j.ijar.2013.04.003

[46] 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 Conference on Knowledge Discovery and Data Mining (KDD) (2016)

[47] Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020). https://doi.org/10.1038/s42256-020-00257-z

[48] DeGrave, A.J., Janizek, J.D., Lee, S.-I.: AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence **3**(7), 610–619 (2021). https://doi.org/10.1038/s42256-021-00338-7

[49] Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A.I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J.R., Teng, Z., Gkrania-Klotsas, E., AIX-COVNET, Rudd,

J.H.F., Sala, E., Schönlieb, C.-B.: Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence **3**(3), 199–217 (2021). https://doi.org/10.1038/s42256-021-00307-0

[50] Ko, J., Baldassano, S.N., Loh, P.-L., Kording, K., Litt, B., Issadore, D.: Machine learning to detect signatures of disease in liquid biopsies – a user's guide. Lab Chip **18**, 395–405 (2018). https://doi.org/10.1039/C7LC00955K

[51] Wang, D., Li, J.-R., Zhang, Y.-H., Chen, L., Huang, T., Cai, Y.-D.: Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms. Genes **9**(3), 155 (2018)

[52] Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D.E., Zou, J.: How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nature Medicine **27**(4), 582–584 (2021). https://doi.org/10.1038/s41591-021-01312-x

[53] Rajkomar, A., Hardt, M., Howell, M.D., Corrado, G., Chin, M.H.: Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine **169**(12), 866–872 (2018). https://doi.org/10.7326/M18-1990

[54] Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D.C.M., Ezer, D., Haert, F.C.v.d., Mugisha, F., Abila, G., Arai, H., Almiraat, H., Proskurnia, J., Snyder, K., Otake-Matsuura, M., Othman, M., Glasmachers, T., Wever, W.d., Teh, Y.W., Khan, M.E., Winne, R.D., Schaul, T., Clopath, C.: Ai for social good: unlocking the opportunity for positive impact. Nature Communications **11**(1), 2468 (2020). https://doi.org/10.1038/s41467-020-15871-z

[55] Kattan, M.W., Hess, K.R., Amin, M.B., Lu, Y., Moons, K.G.M., Gershenwald, J.E., Gimotty, P.A., Guinney, J.H., Halabi, S., Lazar, A.J., Mahar, A.L., Patel, T., Sargent, D.J., Weiser, M.R., Compton, C., members of the AJCC Precision Medicine Core: American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine. CA: A Cancer Journal for Clinicians **66**(5), 370–374 (2016). https://doi.org/10.3322/caac.21339

[56] Alaa, A.M., Gurdasani, D., Harris, A.L., Rashbass, J., van der Schaar, M.: Machine learning to guide the use of adjuvant therapies for breast cancer. Nature Machine Intelligence **3**(8), 716–726 (2021). https://doi.org/10.1038/s42256-021-00353-8

[57] Van der Velden, B.H., Kuijf, H.J., Gilhuijs, K.G., Viergever, M.A.: Explainable artificial intelligence (xai) in deep learning-based medical image analysis. Medical Image Analysis, 102470 (2022)

[58] Rajpurkar, P., O'Connell, C., Schechter, A., Asnani, N., Li, J., Kiani, A., Ball, R.L., Mendelson, M., Maartens, G., van Hoving, D.J., Griesel, R., Ng, A.Y., Boyles, T.H., Lungren, M.P.: Chexaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with hiv. npj Digital Medicine **3**(1), 115 (2020). https://doi.org/10.1038/s41746-020-00322-2

[59] Rudin, C.: Why black box machine learning should be avoided for high-stakes decisions, in brief. Nature Reviews Methods Primers **2**(1), 81 (2022). https://doi.org/10.1038/s43586-022-00172-0

[60] Rudin, C., Wang, C., Coker, B.: The age of secrecy and unfairness in recidivism prediction. Harvard Data Science Review **2**(1) (2020). https://hdsr.mitpress.mit.edu/pub/7z10o269

[61] Ghassemi, M., Oakden-Rayner, L., Beam, A.L.: The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health **3**(11), 745–750 (2021). https://doi.org/10.1016/S2589-7500(21)00208-9

[62] Reyes, M., Meier, R., Pereira, S., Silva, C.A., Dahlweid, F.-M., Tengg-Kobligk, H.v., Summers, R.M., Wiest, R.: On the interpretability of artificial intelligence in radiology: Challenges and opportunities. Radiology: Artificial Intelligence **2**(3), 190043 (2020). https://doi.org/10.1148/ryai.2020190043

[63] Reddy, S.: Explainability and artificial intelligence in medicine. The Lancet Digital Health **4**(4), 214–215 (2022). https://doi.org/10.1016/S2589-7500(22)00029-2

[64] Arcadu, F., Benmansour, F., Maunz, A., Willis, J., Haskova, Z., Prunotto, M.: Deep learning algorithm predicts diabetic retinopathy progression in individual patients. npj Digital Medicine **2**(1), 92 (2019). https://doi.org/10.1038/s41746-019-0172-3

[65] Pierson, E., Cutler, D.M., Leskovec, J., Mullainathan, S., Obermeyer, Z.: An algorithmic approach to reducing unexplained pain disparities in underserved populations. Nature Medicine **27**(1), 136–140 (2021). https://doi.org/10.1038/s41591-020-01192-7

[66] van der Schaar, M., Maxfield, N.: Making machine learning interpretable: a dialog with clinicians. https://www.vanderschaar-lab.com/making-machine-learning-interpretable-a-dialog-with-clinicians/. Accessed: 2023-01-25