Canadian Journal of Cardiology 38 (2022) 204–213

Review

Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology

Jeremy Petch, PhD, MA, BA(H),[a,b,c,d] Shuang Di, MSc, BSc,[a,e,‡] and Walter Nelson, BSc(H)[a,f,‡]

[a] Centre for Data Science and Digital Health, Hamilton Health Sciences, Hamilton, Ontario, Canada

[b] Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada

[c] Division of Cardiology, Department of Medicine, McMaster University, Hamilton, Ontario, Canada

[d] Population Health Research Institute, Hamilton, Ontario, Canada

[e] Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

[f] Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada

ABSTRACT

Many clinicians remain wary of machine learning because of long-standing concerns about "black box" models. "Black box" is shorthand for models that are sufficiently complex that they are not straightforwardly interpretable to humans. Lack of interpretability in predictive models can undermine trust in those models, especially in health care, in which so many decisions are— literally—life and death issues. There has been a recent explosion of research in the field of explainable machine learning aimed at addressing these concerns. The promise of explainable machine learning is considerable, but it is

RÉSUMÉ

De nombreux cliniciens restent méfiants envers l'apprentissage automatique en raison de préoccupations de longue date concernant les modèles à « boîte noire ». Le terme « boîte noire » sert à désigner des modèles suffisamment complexes pour échapper à une interprétation simple par un humain. Le manque d'interprétabilité des modèles prédictifs peut miner la confiance en ces modèles, en particulier dans le domaine des soins de santé, où tant de décisions sont littéralement des questions de vie ou de mort. Il y a eu récemment une explosion de la recherche consacrée à l'apprentissage automatique explicable

Scientific studies employing machine learning (ML) are becoming increasingly common in cardiology.[1-3] And although most Food and Drug Administration- and Health Canada-approved ML algorithms are focused on the brain, lung, or breast, as of the time of writing there are at least 12 approved algorithms with application in cardiology.[4] Yet, despite this proliferation, many clinicians remain wary of ML because of concerns about the "black-box" nature of many ML models that date back to some of the early applications of artificial neural networks to medicine in the 1990s.[5] Concerns about black-box algorithms are routinely highlighted in commentaries in prominent medical journals[6] and are now acknowledged by many ML scientists as one of the primary barriers to adoption of ML in medicine.[7]

The term "black box" is shorthand for models that are sufficiently complex that they are not straightforwardly interpretable to humans. This contrasts with models that are routinely used in medical research, such as linear and logistic regression, in which humans can refer to model coefficients to interpret the model and its predictions. Although not all ML algorithms are uninterpretable (more on this will follow), most ML algorithms producing state-of-the-art results—including deep learning and ensemble models—do suffer from this limitation.

Lack of interpretability in predictive models can undermine trust in those models, especially in health care, in which so many decisions are—literally—life and death issues.[8] No model is perfect, so it is entirely reasonable that a clinician would be weary of blindly trusting the prediction of a model that cannot provide any insight as to the cause of its decision. These concerns have only increased as empirical evidence has begun to accumulate of how ML can encode existing racial bias found in observational health care data sets,[9] thus perpetuating systematic racism through black-box automation.[10]

Black boxes also limit the clinical actionability of model predictions, which further undermines their usefulness to clinicians.[11] A common application of ML in health care is for early warning systems to predict clinical deterioration.[12] But if such systems only warn clinicians that there is a risk of

Received for publication July 30, 2021. Accepted September 8, 2021.

‡These authors contributed equally to this article.

Corresponding author: Dr Jeremy Petch, 90 Highland Park Drive, Dundas, Ontario L9H 6G8, Canada. Tel.: +1-416-476-9039.

E-mail: Jeremy.petch@utoronto.ca

See page 211 for disclosure information.

https://doi.org/10.1016/j.cjca.2021.09.004

important for cardiologists who may encounter these techniques in clinical decision-support tools or novel research papers to have critical understanding of both their strengths and their limitations. This paper reviews key concepts and techniques in the field of explainable machine learning as they apply to cardiology. Key concepts reviewed include interpretability vs explainability and global vs local explanations. Techniques demonstrated include permutation importance, surrogate decision trees, local interpretable model-agnostic explanations, and partial dependence plots. We discuss several limitations with explainability techniques, focusing on the how the nature of explanations as approximations may omit important information about how black-box models work and why they make certain predictions. We conclude by proposing a rule of thumb about when it is appropriate to use black- box models with explanations rather than interpretable models.

visant à répondre à ces préoccupations. L'apprentissage automatique explicable est très prometteur, mais il importe que les cardiologues aient une compréhension critique de ses forces et de ses limites puisqu'ils pourraient le retrouver dans des outils d'aide à la décision clinique ou des rapports de recherche de pointe. Cet article passe en revue les concepts et éléments techniques clés de l'apprentissage automatique explicable, tels qu'ils s'appliquent à la cardiologie. Les concepts clés que nous examinons comprennent l'interprétabilité par rapport à l'explicabilité et les explications globales par rapport aux explications locales. Au nombre des éléments techniques présentés figurent la mesure d'importance des variables à partir de permutations, les arbres de décision de substitution, l'algorithme LIME (*local interpretable model-agnostic explanations*) et les tracés de dépendance partielle. Nous abordons plusieurs limites des éléments techniques d'explicabilité, en attirant l'attention sur le fait que la nature approximative des explications peut mener à l'omission d'informations importantes sur la façon dont les modèles à boîte noire fonctionnent et les fondements de certaines prévisions obtenues à l'aide de ces modèles. Pour conclure, nous proposons une méthode empirique permettant de déterminer quand il est approprié d'utiliser des modèles à boîte noire assortis d'explications plutôt que des modèles interprétables.

deterioration without indicating why, the reason for the alert may not be clear without further assessment, which can delay treatment in time-sensitive situations or waste valuable clinician time in the case of false positives.[13]

In response to the significant challenges posed by black-box models, there has been an explosion of research in recent years in the field of explainable ML.[14] Much of this research is focused not on making black-box models inherently interpretable but, instead, on creating intelligible explanations of how a model works and why it makes specific individual predictions (more on the distinction between interpretability vs explainability follows). Many of these explanations are focused on identifying the variables most driving model predictions or on translating the workings of a model into a transparent design that more closely mirrors evidence-based clinical reasoning (such as a decision tree). Both of these approaches have been shown to be important in enhancing clinician confidence in ML models.[15,16] The promise of explainable ML is considerable: the opportunity to benefit from the state-of-the-art predictive power of techniques such as deep learning with none of the drawbacks of black boxes. However, with this promise comes some important limitations.

This paper reviews key concepts and techniques in the field of explainable ML as they apply to cardiology. Our goal is to provide cardiologists and cardiovascular researchers with a critical understanding of both the benefits and the limitations of explainable ML techniques so they can be informed consumers whether they encounter these techniques embedded within a clinical decision-support tool or in the findings of a novel cardiology study.

Key Concepts in Explainable Machine Learning

Interpretability vs explainability

Although it is common in the literature to see the terms "interpretability" and "explainability" used interchangeably,[17-19] they are distinct concepts, and their conflation can cause significant confusion.[20]

Although there is slight variability in the precise definition of the term "interpretable" in the literature,[21-24] throughout this review we use the term to refer to models in which humans can directly understand how a model operates and the causes of its decisions.[25] Logistic-regression models are interpretable because a human can refer to the weights and odds ratios to understand how the model operates and can refer to coefficients to understand the cause of individual predictions. It should be noted that interpretability is at least somewhat subjective, as it can require expert knowledge of statistics or a domain (such as cardiology) to interpret a model's decisions effectively.[8]

There is a common misperception that all ML models are black boxes, but, in fact, many are inherently interpretable. Examples include decision trees[22] and if-then rule sets generated by algorithms such as **C**ertifiably **O**ptimal **R**u**l**e **L**ist**s** (CORELS).[26] However, as illustrated in Figure 1, some of these inherently interpretable models quickly become less interpretable as their complexity increases. This is true of those models currently producing state-of-the-art results, including deep learning and ensemble methods (the family of algorithms that includes random forest and gradient boosting). Although based on entirely interpretable mathematic functions,[27,28] these models are rendered black boxes by the complexity and scale of their structures. These methods introduce the need for explanations.

Explainability in ML seeks to imbue humans with a high level of understanding of how a model works and reaches decisions without trying to account for all the minutia of its calculations.[29] Conceptually, explainable ML is not dissimilar from how a clinician must distill extremely complex reasoning based on decades of medical education and clinical experience into a plain-language explanation that is understandable to a patient. The clinician's explanation will gloss over many details of their reasoning and may

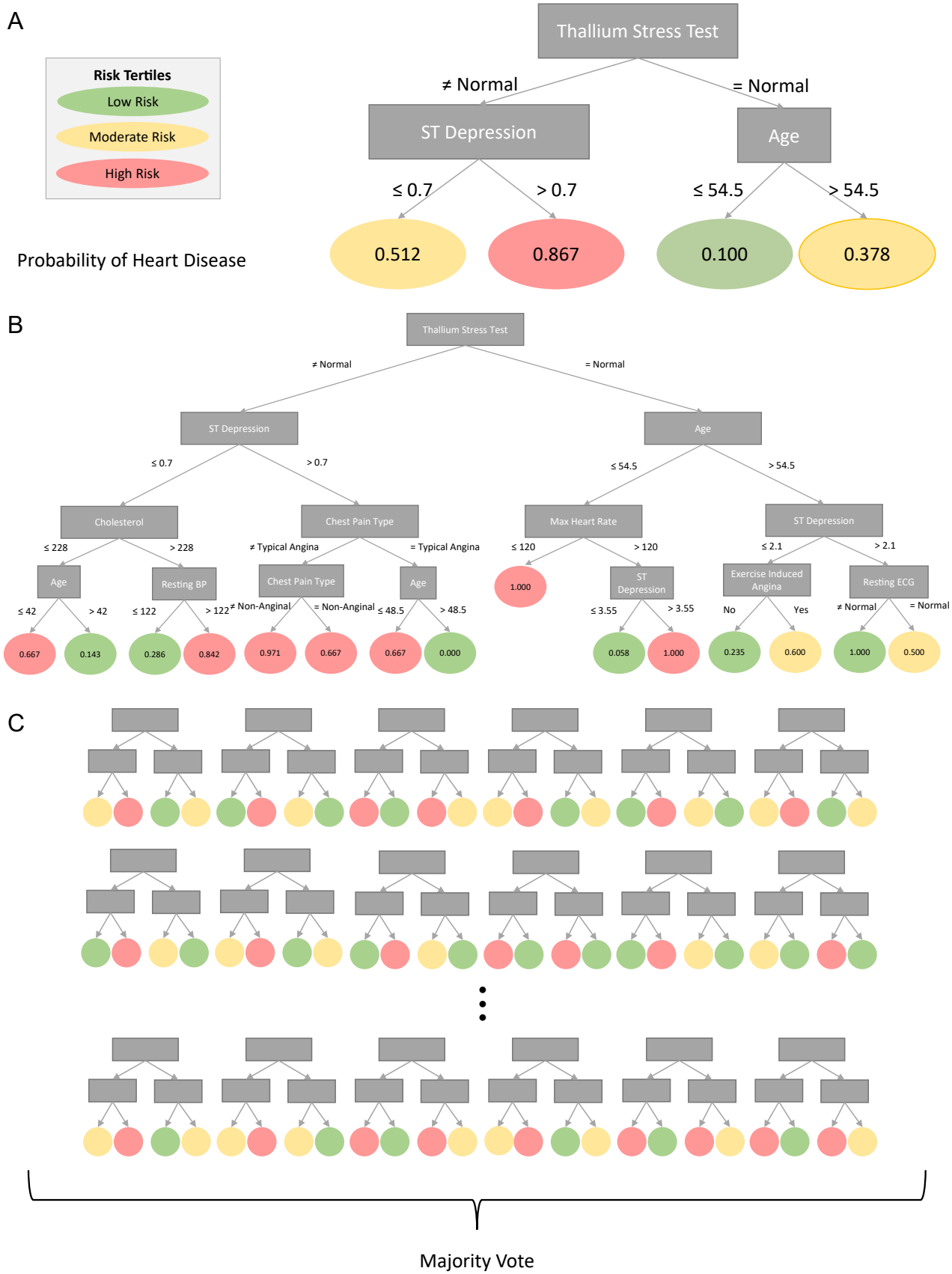

Figure 1. Inherently interpretable models can become black boxes as their complexity and scale increase. Short decision trees (**A**) are easily interpreted, but the ease of interpretability decreases as they grow larger (**B**). Random forests (**C**), which compute hundreds or thousands of individual decision trees and make predictions based on majority vote, are completely uninterpretable.

oversimplify some concepts but, nonetheless, provides a patient with sufficient understanding to be able to provide informed consent. Explainability techniques in ML seek to provide an analogous level of understanding in its users, providing them with sufficient information to make informed decisions about whether they wish to follow a model's recommendation.[7] In practical terms, explainability techniques most often generate a new interpretable model that mostly replicates the behaviour of a black-box model or produce some summary statistics or a visualization to illustrate roughly how inputs are being used by the model.[25] Thus, when interpretations are direct accountings of a model's structure and behaviour, explanations are intelligible approximations of the same.[30]

Global vs local explanations

Explanations for black-box models may either be at the global level (explaining the overall workings of a model) or the local level (explaining how the model reached a particular decision).[31]

Global explainability is critical during model development, when it is used to validate that a model is learning correctly: for example, that the variables driving model predictions are coherent with clinical knowledge.[32] Global techniques are also used for other quality control purposes, such as to assess for biases in training data that can have impact on the accuracy, fairness, or generalizability of a model.[33] There is also growing interest in using global explanations to generate novel hypotheses for scientific study. These approaches leverage the ability of ML to model nonlinear relationships and interactions in data with a very large number of variables, then use explainability techniques extract novel hypotheses that can be tested further.[34-36]

Local explainability aims to provide insight about why a model made a specific prediction, and it has been identified by clinicians as critical in the context of using a model's outputs to inform clinical practice.[16] Local explanations can be incorporated directly into clinical decision support tools in electronic health records to enhance the transparency and the actionability of predictions.[37]

Key Techniques in Explainable Machine Learning

In this section, we discuss the most common types of explainable ML found in the cardiovascular literature. To illustrate how some of these techniques work in practice, we also provide concrete examples using a black-box random-forest[38] classification model that we trained for this review article on the Cleveland Heart Disease Data Set.[39] We employed this data set to enhance the interpretability of our figures for a cardiology readership, but we caution that these figures are for illustrative purposes only; no conclusions about generalizable predictors of heart disease should be drawn from these examples. As black box models are increasingly being applied to medical images, we also trained a convolutional neural network on retinal fundus images from the EyePACS dataset[40,41] to demonstrate how explainability techniques work in these applications.

Variable importance methods

Variable importance methods (called "feature importance" in the ML field) generate explanations by calculating the statistical contribution of each variable to a model's performance. Variable importance methods are routinely used during model development to evaluate whether a model is learning correctly and whether the variables contributing to predictions are clinically plausible. Variable importance methods have been used in the cardiovascular literature for evaluating predictors in applications such as in-hospital length of stay,[42] detection of ventricular arrhythmia,[43] in-stent restenosis,[44] and prediction of risk for cardiovascular disease.[45] Variable importance methods provide global explanations. Variable important methods include "Permutation Importance,"[46] "Mean Decrease in Impurity,"[47] and "Conditional Variable Importance."[48]

To illustrate how these methods work in practice, we generated an example using permutation importance (Fig. 2). Permutation importance measures the decrease in a black-box model's predictive performance when a single variable is randomly shuffled. Permutation importance works by repeatedly retesting a model but each time with a different variable's values mixed up randomly. The method measures predictive performance when each variable is shuffled. If shuffling a variable's values has no impact on predictive performance, the variable is not making a significant contribution to the model's predictions. However, if randomly shuffling the values of a variable results in a large decrease in predictive performance, it means that variable is more important to the model's predictions.

Permutation importance provides straightforward explanations that are relatively easy to understand. They also have the benefit of being computationally efficient in that they do not involve retraining models over and over. However, permutation importance can produce unreliable results with situations in which a model is using highly correlated variables.[48] Conditional variable importance can be used in such cases, although it has the drawback of being less computationally efficient.

Surrogate methods

Surrogate methods explain a black-box model by developing a new interpretable model, such as a logistic regression or a short decision tree, on the predictions of the black-box model. The intent of using the predictions of the black-box model for training (rather than ground truth), is for the new interpretable model to be as faithful as possible to the black-box model. Users seeking an explanation for the black-box model may then refer to the interpretable model for an approximation of how the black box works. Surrogate methods have been employed in the cardiology literature for applications such as explaining ML-prediction models for hypertension[19] and to explain electrocardiographic classifications.[49,50] Surrogate methods can be both global and local. Global surrogate methods include decision trees[22] and logistic-linear regression. Local surrogate methods include Local Interpretable Model-Agnostic Explanations (LIME)[51] and Shapley Additive Explanations (SHAP).[52] Gradient-Weighted Class Activation Mapping (Grad-CAM),[53] guided

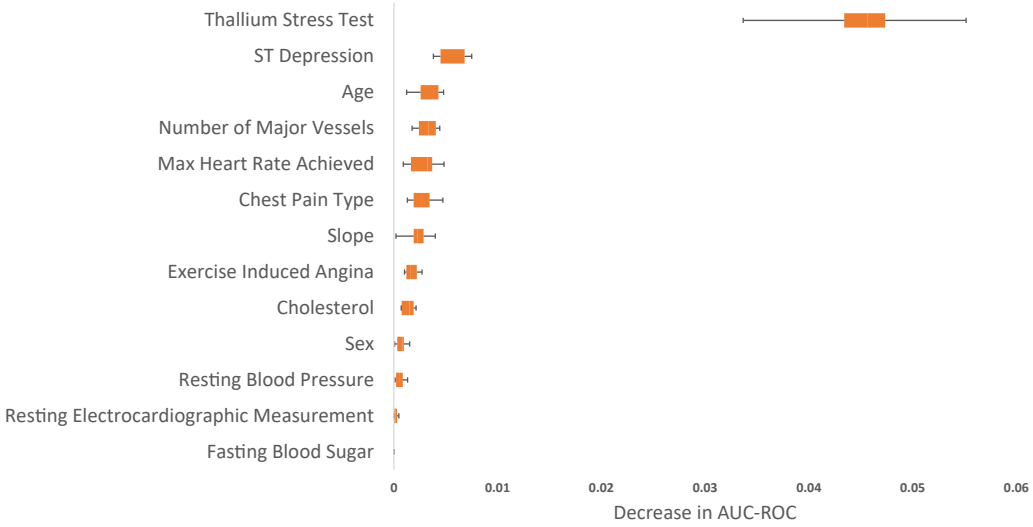

Figure 2. Relative importance of each predictor in the random-forest model for predicting probability of heart disease. This figure illustrates the relative importance of each variable in the random-forest model we trained to predict the probability of a patient having heart disease. The *x*-axis illustrates the decrease in area under the curve—receiver-operating characteristic (AUC-ROC) when a given variable's values are shuffled. This figure indicates that thallium stress test is far and away the most important predictor in our model (AUC drops by 0.05 when the variable is shuffled), whereas measures such as fasting blood sugar and resting electrocardiogram (ECG) are relatively unimportant (negligible decrease in AUC when these variables are shuffled).

back propagation,[54] and integrated gradients[55] are local surrogate methods tailored for image-classification models.

To illustrate how this type of explanation works in practice, we provide 3 examples of the most deployed techniques. The first uses short decision trees to provide both global and local explanations for our random-forest model. The second uses LIME to explain specific predictions of heart disease from the same model. The third uses Grad-CAM to illustrate how local explanations can be applied to image classification models such as our convolutional neural network for predicting retinopathy.

Figure 3 was created by training a single new decision tree on the predictions of our random- forest model. The global tree diagram provides a representation of how the random forest is generally working. Individual predictions can then be analyzed by tracing the path patients took as they moved through the global tree. An advantage of this type of explanation is that it makes it easy to simultaneously understand the overall workings of the model and the reasons for each specific prediction. Decision trees also provide rule-based logic that is analogous to clinical decision rules, making it a good fit for fields such as cardiology.

Figure 4 depicts explanations produced by LIME for 2 local predictions of our heart-failure model. As local explainability in health care is geared toward clinicians, we illustrate these predictions as being embedded as part of a clinical decision support tool in an electronic health record such as Epic (Epic Systems Corporation, Verona, Wisconsin, USA).[37] The advantage of this type of explainability technique is that it provides easily understood explanations of which clinical factors contribute to each prediction. The ability to integrate this type of explanation into an electronic health record is also notable, as it could improve the actionability of a black-box model by seamlessly integrating predictions and easily understood explanations into clinical workflow.

In Figure 5, we illustrate how surrogate methods can be used to provide local explanations for image classification. These methods provide explanations by indicating which areas of an image are factors in a prediction, much as LIME in the previous example identified contributing clinical factors. In practice these explanations take the form of highlighted pixels in the image that are associated with a specific classification prediction. This can increase confidence in a model by allowing a clinician to verify that it is using the correct part of an image in its prediction, as in the case of the highlighted neovascularization of the disc in Figure 5.

Visualization methods

Visualization methods for black-box explanations can be used to illustrate relationships between an outcome and a set of variables of interest. This family of techniques has been used in the cardiology literature to provide insight into the relative importance of cardiovascular risk factors in patients with diabetes.[34] Visualization methods provide global explanations for black-box models. Visualization methods include partial dependence plots (PDPs),[56] centered individual conditional expectation (ICE) plots,[29] and accumulated local effects (ALE) plots.[57]

To illustrate how this type of explanation works in practice, we provide an example using a PDP. PDPs illustrate how the predicted outcome of a model changes as a set of independent variables change (the marginal effect, in statistical terms). In Figure 6, we use a PDP to illustrate the effect of age and cholesterol on the probability of having heart disease in our random forest model. By incorporating 2 independent variables in the same plot, we can illustrate the interaction between these variables that is captured by the model.

An important advantage of PDPs over some other explainability techniques is that they can illustrate the

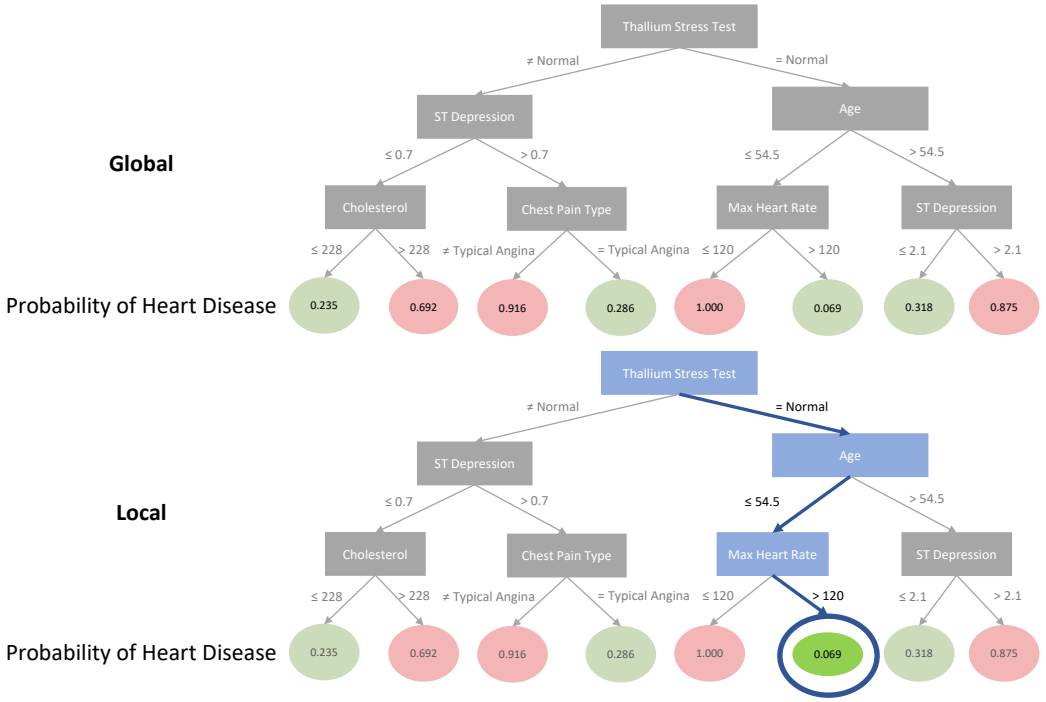

Figure 3. Global and local decision trees to explain a random forest model for predicting heart disease. Decision trees are read from the top down, so the global diagram shows that the model is first analyzing whether patients have normal results on their thallium stress tests. If the thallium stress test shows a defect, the model looks at the patient's ST depression, and so on. The local diagram illustrates the reason for an individual prediction by highlighting the path a patient took down the tree. In this case, a patient was predicted to have a very low probability of heart disease because there was a normal result for the thallium stress test, was less than 54.5 years old, and had a maximum heart rate of greater than 120.

relationship between variables and predictions, even when those relationships are nonlinear. And although the 3-dimensional plots take some orientation for new users, they provide a relatively simple and intuitive illustration of how the average prediction of a model changes when dependent variables change. However, as humans can only perceive visualizations in up to 3 dimensions, the maximum number of dependent variables that can be plotted simultaneously is 2, which may obscure particularly complex interactions between sets of variables. PDPs also assume that the variables shown in the plot are not correlated with other variables used in the model, and they are not well suited to illustrating heterogenous effects[56] (heterogenous effects can be modelled using ICE plots, but these plots have their own tradeoffs, as they can only illustrate the effect of a single independent variable).[29]

Discussion

Explainability techniques seek to address concerns about black-box ML models that have long stood as a significant barrier to adoption of ML in health care. Many of these techniques target clinicians' needs for transparency,[16] and there is some empirical evidence that they can increase clinician trust in black-box models.[15] However, the landscape of explainable ML is evolving rapidly, and current explainability techniques suffer from limitations that cardiologists should understand before making use of them in a clinical or scientific setting. In this section, we discuss these limitations and propose a rule of thumb for when the use of black-box models with explanations is justified vs when they should be avoided in favour of interpretable models.

Limitations of explainability techniques

The most notable limitation of explainability techniques is that most of them are approximations of black-box models and therefore do not precisely account for the inner workings of those models. For example, the global decision tree explanation depicted in Figure 3 uses a single tree to explain the workings of a random-forest model composed of 100 individual decision trees, each of which was trained on a subset of the variables in the data set. Further, to enhance the interpretability of the explanation, we elected to generate a short decision tree, only 3 variables deep. Thus, there is no question that this explanation provides only a post hoc approximation of the model's functioning, and therefore that much of the complexity of the model's operation remains unaccounted for. This led Babic et al. to worry that explanations offer only an "ersatz understanding" of black-box models.[30] Although this limitation is important to acknowledge, it is arguably true of all forms of explanation of complex phenomena, including our earlier analogy to the explanations clinicians provide patients. These often condense decades of education and experience into short narratives that do not fully account for the totality of clinicians' reasoning yet are understood to be sufficient for patients to provide informed consent to recommended courses of treatment.

A key advantage of many ML methodologies is that they can model nonlinear relationships, but in these cases the

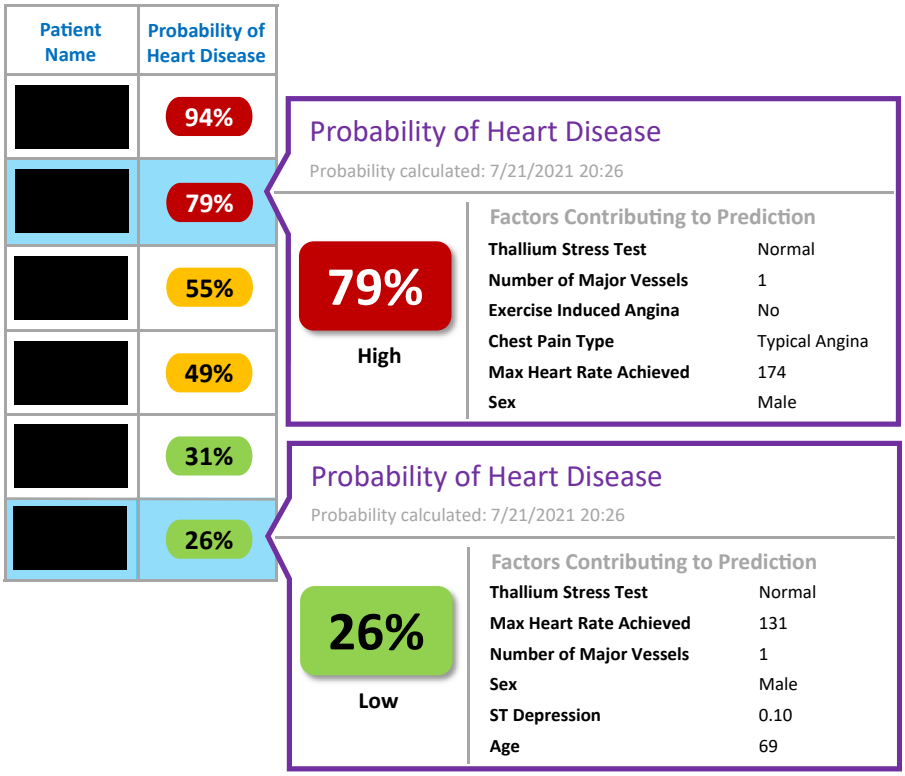

Figure 4. Explanation for predictions of heart disease generated using Local Interpretable Model-Agnostic Explanations (LIME). This figure demonstrates how a local explanation using the LIME algorithm can be embedded as clinical decision support in an electronic health record such as Epic. Probabilities are color coded to draw clinicians' attention to patients predicted to be at high risk of heart disease. The clinical factors of greatest importance to the prediction are displayed on the right to enhance the transparency and actionability of the prediction for clinicians.

strategy of explaining black-box models through approximations may be particularly limiting. This is because many explainability techniques use linear models, which may significantly misrepresent black-box models that are highly nonlinear.[20] Even with nonlinear explainability techniques such as decision trees, the relative simplicity of the explanations compared with the black-box models means that any nonlinear relationships surfaced through the explanation are likely to be oversimplifications and thus should be interpreted with caution. It should be noted that this limitation does not affect every explainability technique, only those that generate explanations using a secondary post hoc model. For example, visualization methods such as PDPs (Fig. 6) produce much more accurate representations of nonlinear relationships captured by black-box models.[56]

A related limitation of black box explanations is that they are not always entirely illuminating or actionable. For example, Figure 4 illustrates 2 individual predictions, 1 high risk, and the other low. Yet Thallium Stress Test = Normal, Number of Major Vessels = 1, and Sex = Male are listed as factors contributing to both predictions. Even if these explanations are entirely correct (and not the product of a poor

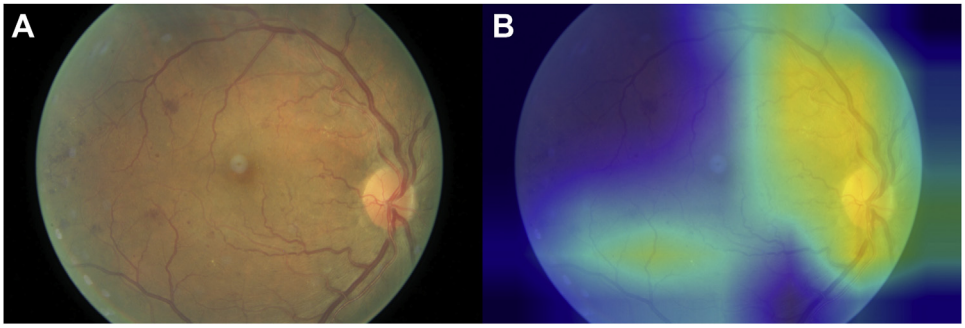

Figure 5. Explanation of a convolutional neural network's prediction of proliferative diabetic retinopathy from a retinal fundus photograph using Gradient-Weighted Class Activation Mapping (Grad-CAM). The photo on the left (**A**) was corrected classified by our model as having proliferative diabetic retinopathy. The photo on the right (**B**) is the explanation from Grad-CAM; it is the same image but with the pixels most contributing to the prediction shaded **yellow**, and those contributing least shaded **blue**.

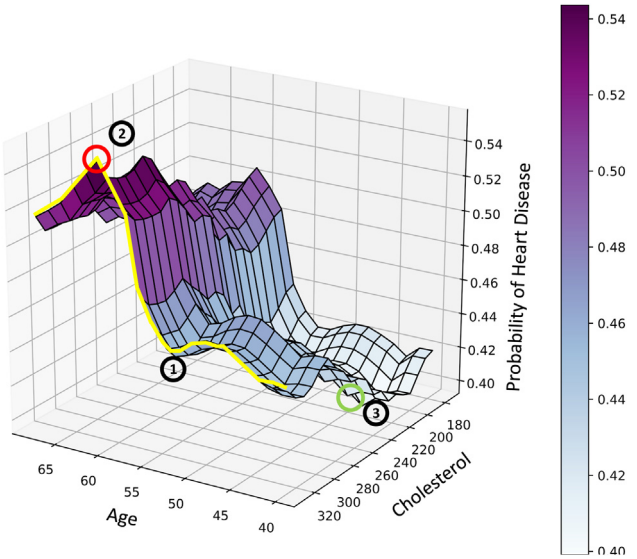

Figure 6. The effect of age and cholesterol on the probability of heart disease. This figure illustrates the marginal effect of age and cholesterol on the probability of having heart disease in a random-forest model. We have made several annotations to acclimatize readers who are new to this type of plot: the **yellow line** (1) shows the nonlinear effect of age on the probability of developing heart disease, in the subgroup of patients with high total cholesterol level (> 320); the **red circle** (2) indicates older patients with high cholesterol levels have a relatively higher chance of developing heart disease; and the **green circle** (3) illustrates that, on average, young patients with low cholesterol levels are less likely to develop heart disease.

approximation), their inclusion in a clinical decision-support tool may be of limited value to a busy clinician, at least not without significantly more information about the interactions among these and the other factors. Similarly, the image classification explanations illustrated in Figure 5 tell us only what data the model is using, not how it is using it. This can undermine their value as explanations with instances in which the maps for different classifications are visually indistinguishable,[20] although the magnitude of this concern in the context of medical imaging has not yet been clearly established in the literature.

We should also note that explainability techniques for ML have the same limitations as other statistical techniques when applied to observational data; they can only identify correlation, not causation. Although it is tempting to assume variables with high predictive power are risk factors with some mechanistic role in the pathogenesis of a disease, such determinations can only be made reliably in instances that do not suffer from confounding and selection bias.[58] This does not mean that such techniques have no potential role in the discovery of novel risk factors, only that explanations from black-box models trained on observational data sets should be restricted to hypothesis generation and followed up with studies appropriately designed to test these hypotheses.

When is the use of black-box model explanations justified?

The limitations of explainable ML have led Rudin to argue that the use of black box models should be avoided in favour of interpretable models when making high-stake decisions.[20] Rudin points out that the current proliferation of black box models is based on a widespread assumption that there is always a tradeoff between a model's accuracy and its interpretability. Yet, although deep learning is currently outperforming other types of models on unstructured data such as free text[59] and images,[60] it is not uncommon to see interpretable methods performing on par with black-box models on structured clinical data.[61] Given these considerations, when is it justified to employ a black-box model with explanations rather than an interpretable model?

We propose a rule of thumb for when the use of black-box models (with accompanying explanations) may be appropriate. From the outset, data scientists should train models using both interpretable and black-box methods to assess whether there is, in fact, an accuracy vs interpretability tradeoff in the specific case on which they are working. If there is no meaningful difference in accuracy between an interpretable model and a black box, an interpretable method should be used. However, if a black-box model does provide a higher degree of accuracy, the stakes of the decision should be considered. If the decision that will be informed by the model is relatively low stake, a small improvement in accuracy may justify the use of a black box. However, if the stakes are high, it is reasonable to require a greater improvement in accuracy before sacrificing interpretability. Ideally, gains in accuracy from black-box methods should be sufficient to translate into meaningful improvements in clinical outcomes such as reduced morbidity or mortality. If the use of a black box model can be justified, explainability techniques should be employed to make the model and its predictions as transparent as possible, but clinicians should be aware of their limitations and be cautious of overinterpreting, which can lead to narrative fallacies.[30]

Conclusions

Like the broader field of ML, explainable ML is developing rapidly. Although explainability techniques hold the promise to provide all the predictive power of black-box models without the drawbacks, current techniques have limitations that are important for cardiologists and cardiovascular research to appreciate, so they can make informed decisions about when and how to use them.

Funding Sources

This review was funded from the core-operating budget of the Centre for Data Science and Digital Health (CREATE) at Hamilton Health Sciences. The lead author is the Director of CREATE. The larger organization of Hamilton Health Sciences had no role in the design, conduct or content of the review.

Disclosures

The authors have no conflicts of interest to disclose.

References

1. Quer G, Arnaout R, Henne M, Arnaout R. Machine learning and the future of cardiovascular care: JACC State-of-the-Art Review. J Am Coll Cardiol 2021;77:300-13.

2. Krittanawong C, Johnson KW, Rosenson RS, et al. Deep learning for cardiovascular medicine: a practical primer. Eur Heart J 2019;40:2058-2069C.

3. Iannattone PA, Zhao X, VanHouten J, Garg A, Huynh T. Artificial intelligence for diagnosis of acute coronary syndromes: a meta-analysis of machine learning approaches. Can J Cardiol 2020;36:577-83.

4. ACR DSI: FDA cleared AI algorithms. American College of Radiology, Data Science Institute. Available at: https://models.acrdsi.org/. Accessed July 15, 2021.

5. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol 1996;49:1225-31.

6. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA 2017;318:517-8.

7. Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput Appl 2020;32:18069-83.

8. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. Data Mining Knowl Discov 2020;10:1-13.

9. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447-53.

10. Benjamin R. Assessing risk, automating racism. Science 2019;366:421-2.

11. Guidi JL, Clark K, Upton MT, et al. Clinician perception of the effectiveness of an automated early warning and response system for sepsis in an academic medical center. Ann Am Thorac Soc 2015;12:1514-9.

12. Muralitharan S, Nelson W, Di S, et al. Machine learning-based early warning systems for clinical deterioration: systematic scoping review. J Med Internet Res 2021;23. e25187.

13. Umscheid CA, Betesh J, Vanzandbergen C, et al. Development, implementation, and impact of an automated early warning and response system for sepsis. J Hosp Med 2015;10:26-31.

14. Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C. Interpretable machine learning: fundamental principles and 10 grand challenges. Published online 2021. Available at: http://arxiv.org/abs/2103.11251. Accessed July 13, 2021.

15. Lahav O, Mastronarde N, van der Schaar M. What is interpretable? Using machine learning to design interpretable decision-support systems. Published online 2018. Available at: http://arxiv.org/abs/1811.10799. Accessed August 8, 2021.

16. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. In: Doshi-Velez F, Fackler J, Jung K, et al., eds. Proceedings of the 4th Machine Learning for Healthcare Conference106. Proceedings of Machine Learning Research. PMLR, 2019:359-80.

17. Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: a survey on methods and metrics. Electron 2019;8:832.

18. Reyes M, Meier R, Pereira S, et al. On the interpretability of artificial intelligence in radiology: challenges and opportunities. Radiol Artif Intell 2020;2. e190043.

19. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak 2019;19:146.

20. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206-15.

21. Van Lent M, Fisher W, Mancuso M. An explainable artificial intelligence system for small-unit tactical behavior, In: Proceedings of the National Conference on Artificial Intelligence. Menlo Park, CA; Cambridge, MA; London: AAAI Press; MIT Press, 2004. Available at, https://www.aaai.org/Papers/IAAI/2004/IAAI04-019.pdf. Accessed December 16, 2021.

22. Piltaver R, Luštrek M, Gams M, Martiň Cí C-Ipší S. What makes classification trees comprehensible? Expert Syst Appl 2016;62:333-46.

23. Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif Intell 2019;267:1-38.

24. Elshawi R, Sherif Y, Al-Mallah M, Sakr S. Interpretability in healthcare a comparative study of local machine learning interpretability techniques. In: Proceedings IEEE Symposium on Computer-Based Medical Systems2019. Institute of Electrical and Electronics Engineers Inc., 2019:275-80.

25. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. In: Proceedings 2018 IEEE 5th International Conference on Data Science and Advanced Analytics. DSAA 2018;2019:80-9.

26. Angelino E, Larus-Stone N, Alabi D, Seltzer M, Rudin C. Learning certifiably optimal rule lists for categorical data. J Mach Learn Res 2018;18:1-78.

27. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge, MA: MIT Press, 2016. Available at: https://www.deeplearningbook.org/. Accessed December 16, 2021.

28. Shalev-Shwartz S, Ben-David S. Understanding Machine Learning: From Theory to Algorithms. Cambridge, UK: Cambridge University Press. Available at: http://www.cs.huji.ac.il/∼shais/UnderstandingMachineLearning. Accessed December 16, 2021.

29. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat 2015;24:44-65.

30. Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. Science 2021;373:284-6.

31. Molnar C. Interpretable machine learning: a guide for making black box models explainable. Available at: https://christophm.github.io/interpretable-ml-book. Accessed August 20, 2021.

32. 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 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015;1721−1730.

33. Pereira S, Meier R, Mckinley R, et al. Enhancing interpretability of automatically extracted machine learning features: application to a RBM-random forest system on brain lesion segmentation. Med Image Anal 2018;44:228-44.

34. Rawshani A, Rawshani A, Sattar N, et al. Relative prognostic importance and optimal levels of risk factors for mortality and cardiovascular outcomes in type 1 diabetes mellitus. Circulation 2019;139:1900-12.

35. Al-Dury N, Ravn-Fischer A, Hollenberg J, et al. Identifying the relative importance of predictors of survival in out of hospital cardiac arrest: a machine learning study. Scand J Trauma Resusc Emerg Med 2020;28:60.

36. Zhu F, Ju Y, Wang W, et al. Metagenome-wide association of gut microbiome features for schizophrenia. Nat Commun 2020;11:1612.

37. Razavian N, Major VJ, Sudarshan M, et al. A validated, real-time prediction model for favorable outcomes in hospitalized COVID-19 patients. NPJ Digit Med 2020;3:130.

38. Breiman L. Random forests. Mach Learn 2001;45:5-32.

39. Janosi A, Steinbrunn W, Pfisterer M, Detrano R. UCI machine learning repository: heart disease data set. Published 2019, https://archive.ics.uci.edu/ml/datasets/Heart+Disease. Accessed July 27, 2021.

40. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402-10.

41. Cuadros J, Bresnick G. EyePACS: an adaptable telemedicine system for diabetic retinopathy screening. J Diabetes Sci Technol 2009;3:509.

42. Daghistani TA, Elshawi R, Sakr S, Ahmed AM, Al-Thwayee A, Al-Mallah MH. Predictors of in-hospital length of stay among cardiac patients: a machine learning approach. Int J Cardiol 2019;288:140-7.

43. Bhattacharya M, Lu DY, Kudchadkar SM, et al. Identifying ventricular arrhythmias and their predictors by applying machine learning methods to electronic health records in patients with hypertrophic cardiomyopathy (HCM-VAr-Risk Model). Am J Cardiol 2019;123:1681-9.

44. Avram R, Olgin JE, Tison GH. The rise of open-sourced machine learning in small and imbalanced datasets: predicting in-stent restenosis. Can J Cardiol 2020;36:1574-6.

45. Alaa AM, Bolton T, Angelantonio E Di, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. PLoS One 2019;14. e0213653.

46. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics 2010;26:1340-7.

47. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Oxfordshire, UK: Routledge, 1984.

48. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics 2008;9:307.

49. Jones MY, Deligianni F, Dalton J. Improving ECG classification interpretability using saliency maps. In: Proceedings IEEE 20th International Conference on Bioinformatics and Bioengineering, BIBE 2020;2020:675−682.

50. Hicks SA, Isaksen JL, Thambawita V, et al. Explaining deep neural networks for knowledge discovery in electrocardiogram analysis. Sci Rep 2021;11:1-11.

51. Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining13-17. ACM, 2016:1135-44.

52. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 2017:4766-75.

53. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2020;128:336-59.

54. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: the all convolutional net. In: Third International Conference on Learning Representations. ICLR 2015, Workshop Track Proceedings 2015:1-14.

55. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: 34th International Conference on Machine Learning. ICML 2017;2017(Vol 7):5109-18.

56. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189-232.

57. Apley DW, Zhu J. Visualizing the effects of predictor variables in black box supervised learning models. J R Stat Soc Ser B Stat Methodol 2020;82:1059-86.

58. Schooling CM, Jones HE. Clarifying questions about "risk factors": predictors versus explanation. Emerg Themes Epidemiol 2018;15:1-6.

59. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL HLT 2019-2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies—Proceedings of the Conference. Vol 1. 2019:4171-86. Available at: https://arxiv.org/pdf/1810.04805.pdf. Accessed July 27, 2021.

60. Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med 2021;4:65.

61. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12-22.