BJR https://doi.org/10.1259/bjr.20230142

Received: 11 February 2023 | Accepted: 26 May 2023 | Published online: 26 July 2023

Cite this article as:
Cui S, Traverso A, Niraula D, Zou J, Luo Y, Owen D, et al. Interpretable artificial intelligence in radiology and radiation oncology. *Br J Radiol* (2023) 10.1259/bjr.20230142.

AI IN IMAGING AND THERAPY: INNOVATIONS, ETHICS, AND IMPACT: REVIEW ARTICLE

Interpretable artificial intelligence in radiology and radiation oncology

[1]SUNAN CUI, PhD, [2]ALBERTO TRAVERSO, PhD, [3]DIPESH NIRAULA, PhD, [4]JIAREN ZOU, MS, [3]YI LUO, PhD, [5]DAWN OWEN, MD, PhD, [3]ISSAM EL NAQA, PhD and [4]LISE WEI, PhD

[1]Department of Radiation Oncology, University of Washington, Seattle, WA, United States
[2]Department of Radiotherapy, Maastro Clinic, Maastricht, Netherlands
[3]Department of Machine Learning, Moffitt Cancer Center, FL, United States
[4]Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, United States
[5]Department of Radiation Oncology, Mayo Clinic, Rochester, MN, United States

Address correspondence to: Lise Wei
E-mail: *liswei@umich.edu*

ABSTRACT

Artificial intelligence has been introduced to clinical practice, especially radiology and radiation oncology, from image segmentation, diagnosis, treatment planning and prognosis. It is not only crucial to have an accurate artificial intelligence model, but also to understand the internal logic and gain the trust of the experts. This review is intended to provide some insights into core concepts of the interpretability, the state-of-the-art methods for understanding the machine learning models, the evaluation of these methods, identifying some challenges and limits of them, and gives some examples of medical applications.

INTRODUCTION

Artificial intelligence (AI) techniques have been widely applied in numerous medical fields, including radiology, radiation oncology.[1] Radiology and radiation oncology are among the frontiers in medicine that are strongly interested in integrating AI into the clinical workflow, due to their heavy usage of different data (multimodality images, treatment dose plans, clinical information) and computational methods. AI methods have the potential to be used from upstream image reconstruction, registration, segmentation, disease diagnosis, and plan optimization, to therapy prognosis and prediction. Several validated AI models have been implemented into the clinical setting in the form of in-house or commercial softwares[2] . In the year 2019 alone, 77 AI-based medical devices were approved by the Food and Drug Administration (FDA) in the USA and 100 were Conformité Européenne (CE)-marked in Europe.[1,3]

Despite the popularity of these models, there are still concerns and even skepticism about the clinical adoption of these advanced methods. One of the main challenges we face is the intrinsic "black box" nature of some of the AI methods, especially for complex deep learning algorithms. Outside of healthcare, many machine learning (ML) and deep learning (DL) systems do not require interpretability, such as tasks where unacceptable results will not cause harm, or the problem is well studied so that we could trust the ML's decision.[4] However, for medical applications, the above two points don't hold. Incorrect or inaccurate results can cause severe consequences for the patient and a lot of the problems are not well-understood. Thus, we believe that it is essential to develop interpretability into AI systems in medicine to build appropriate safeguard mechanisms.

AI and DL have been explored in multiple medical specialties. For physicians, some of the biggest challenges to the adoption of AI algorithms include: (1) the lack of understanding of a ML model prediction; and (2) esoteric data components used in the model training. Many AI models are generated within single institutions with institution specific data that may not be entirely applicable to another patient population. This speaks to possible bias in the data on which various models are based.[5] Enhancing the clarity of AI inputs and outputs will play a pivotal role in the incorporation of ML into medical decision-making processes. For example, more clinicians are accepting the use of radiomics for the prediction of recurrent cancer and for delineating disease that lies beyond simple morphologic features

on clinical imaging.[6,7] The realm of AI has also extended into genomic signatures and multifaceted models for individualizing diagnosis and standard of care recommendations.[8,9]

ML holds great promise in medicine, but work needs to be done to enhance interpretability and generalizability of various models. AI and ML/DL workflows need physician input to ensure clinical applicability and meaningful patient outcomes.

SCOPE AND CATEGORIES OF AI INTERPRETABILITY

In the literature, interpretability and explainability are often used interchangeably.[10] There is no agreement within the ML and DL community on the definition of interpretability and explainability. Rudin describes the AI interpretability as model predictions being inherently interpretable by humans, while AI explainability is to create a second model to explain the first black box model potentially the underlying mechanics leading to a certain prediction.[11] Despite the quasi-scientific characteristic of the definition of interpretable and explainable AI, the objective of this review is to introduce different interpretable/explainable methods, show examples of applying these approaches, especially in the medical applications, discuss ways to evaluate interpretability, and to provide insights to streamline these advances of ML into clinical practice.

Kim categorized AI interpretability into before (explaining data), during (building inherently interpretable models) and after (post-training interpretability methods).[12] We use the same categorization of Pre-Model *vs* In-Model *vs* Post-Model or before (explaining data), during (building inherently interpretable models) and after (post-training interpretability methods) as the main outline of introducing different interpretable methods. There are other categorizations, but the basic idea is similar. For instance, Lipton proposed methods of making AI more interpretable by either solving the transparent box design problem known as transparent model or explaining the black-box problem.[13] A recent paper assembles seven researchers with varying roles and viewpoints in the field to delve into the concept of explainable AI (XAI). It provides a functional definition and conceptual framework or model that can be employed in the context of XAI.[14] A summary of the interpretable methods is shown in Figure 1.

Explaining data (Pre-model)

Explaining data is the first step for a specific task. Analyzing and summarizing a data set can be simply done by statistical graphs or other visualization tools, which is sometimes ignored by the researchers. Analyzing data distributions can improve the model performance and the understanding of the system. However, this becomes cumbersome for high-dimensional data, which can be resolved by using some dimension reduction methods. Principal component analysis (PCA) is a traditional linear technique for dimensionality reduction. It projects data points onto several principal components to obtain a low-dimensional representation of the data that preserves as much variation of the data as possible. The principal components can be shown to be the eigenvectors of the covariance matrix of the data. However, for high-dimensional data that lies on or near a low-dimensional, non-linear manifold, linear techniques may be suboptimal in identifying the underlying structure of the data. Several nonlinear dimensionality reduction techniques have been developed, including Sammon mapping,[15] Stochastic Neighbor Embedding (SNE),[16] Isomap,[17] Locally Linear Embedding (LLE).[18] More recent approaches include t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP).[19–21] Clustering methods have been applied to visualization of single-cell data wherein both methods preserved local data structure, but UMAP revealed more global structure and provided fast run times. Other applications include identification of prognostic tumor subpopulation[22] and assisting clustering of virus mutation dataset.[23] Despite the popularity of t-SNE and UMAP, careful initialization, learning rate and kernel selection are required to accurately preserve global data structure.[20,24]

Building inherently interpretable models (During-model)

Some ML methods are inherently easier to understand or transparent, such as the linear regression,[25] logistic regression,[26]

Figure 1. A taxonomy of interpretable and explainable ML and DL techniques, Leaf node represents simple, well-known examples. This is not a complete set of all techniques. DL, deep learning; ML, machine learning; PCA, principal component analysis; tSNE, t-distributed Stochastic Neighbor Embedding; UMAP, Uniform Manifold Approximation and Projection.

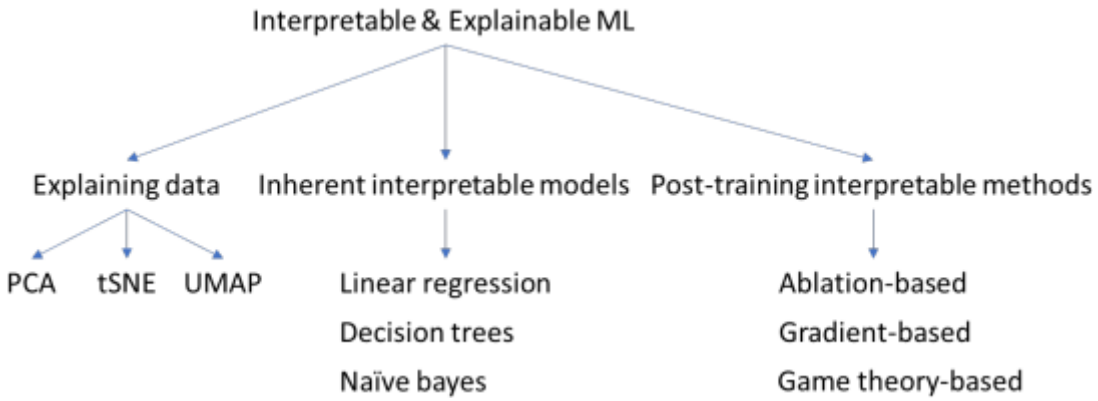
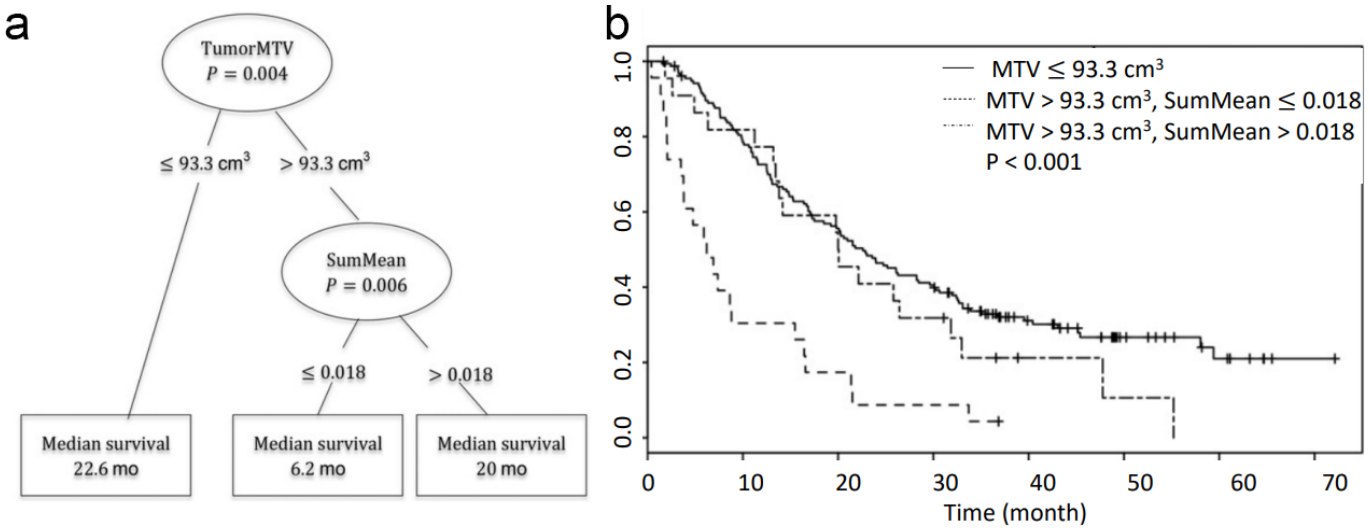


Figure 2. A conditional inference tree for locally advanced NSCLC patients was built. Consideration of both variables jointly yielded an optimal tumor MTV cutpoint of 93.3 cm^3 (P = 0.004) and an optimal SumMean cutpoint of 0.018 (P = 0.006) that was applicable only to patients with large tumor MTV values. Patients with high tumor MTV and low SumMean had significantly inferior OS (median, 6.2 mo) when compared with patients with high tumor MTV and high SumMean (median, 20.0 mo) or patients with low tumor MTV (median, 22.6 mo) (log-rank P < 0.001 across the 3 groups).[31] NSCLC, non-small cell lung cancer; MTV, metabolic tumor volume; OS, overall survival.



decision tree,[27] rulefit,[28] naïve bayes,[29] etc. the goal is to build a model which is interpretable based on the data and task.

Decision trees are non-parametric models for supervised learning tasks that are easily interpretable. It consists of a root node, internal nodes or decision nodes, and leaf nodes or terminal nodes, which can solve both classification and regression problems. Decision tree learning utilizes a divide and conquer approach, carrying out a greedy search to pinpoint the best split point within a tree. This splitting procedure recurs in a top-down manner until all or most samples have been classified. While pruning successfully selects the appropriate tree size, the interpretation of the trees is impacted by skewed variable selection. To enhance interpretability, a unified framework for recursive partitioning has been proposed. This framework integrates tree-structured regression models within a comprehensive theory of conditional inference procedures.[30] This method is used in a work that explored if positron emission tomography (PET) textural features can provide additional prognostic information.[31] Specifically, the conditional inference tree was utilized to identify optimal thresholds for the features (Figure 2).

Post-training interpretability methods (Post-model)

There are several types of post-training methods for interpretability of complex models like DNNs, including ablation test, gradient-based, game theory-based, etc.

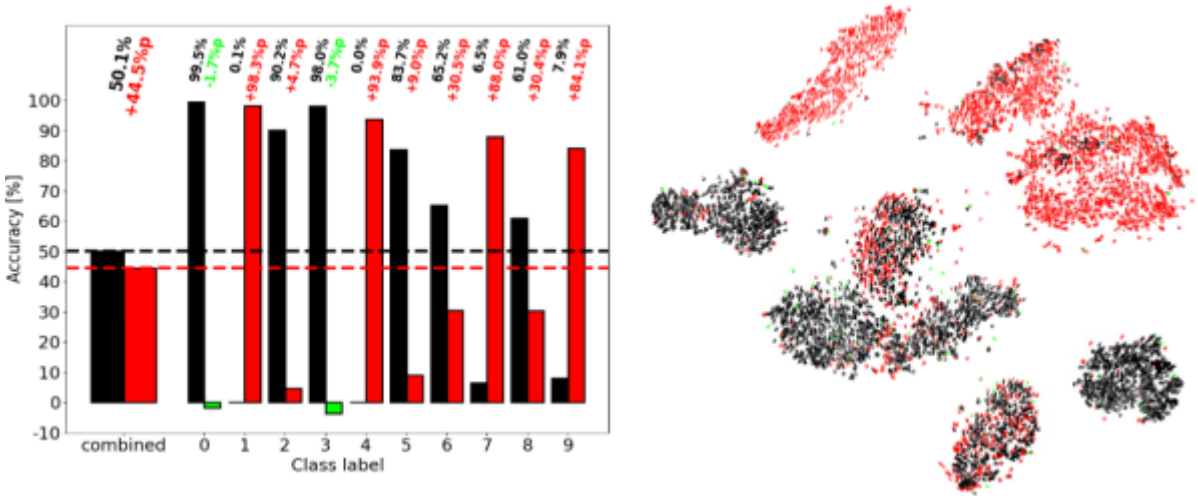
Ablation test:

Certain features (or data points, structure) can be removed and see the impact to the model. Meyes et al did ablation studies in neural networks to investigate the effect of ablating single unit (or pairwise) in a shallow multilayer perceptron (MLP) or VGG network.[32] Figure 3 showed the accuracy change as well as the detailed digit classification performance (visualized by t-SNE) when ablating single unit. Ablation studies by perturbing the data and retraining can be expensive. Koh et al proposed to use influence functions, which is a classic technique from robust statistics[33] that shows how the model parameters change as we change a training point a bit.[34] The idea is trying to see the impact of a training point on a prediction if the values of the training points were changed slightly or if we did not have this training point. As it would be prohibitively expensive to perturb the data and retain the model, they proposed a method to efficiently calculating the influence for either upweighting a training point or perturbing the input. Influence functions could help us understand how the models rely on and extrapolate from the training data by showing the training points that are important for the prediction.

Gradient-based approach:

Ancona et al discussed salience and sensitivity methods within the gradient-based attribution approaches.[35] Sensitivity methods concentrate on illustrating the variations in network output when one or more input features undergo perturbation. On the other hand, salience methods portray the marginal impact of a feature on the output, relative to the same input from which this feature has been eliminated. Deconvolution[36] and guided backpropagation[37] techniques are examples of sensitivity methods. These methods enhance the sensitivity maps and thus the explanation clarity by adjusting the gradients of ReLU functions. They discard negative values during backpropagation computation, which more vividly displays features that instigated activations of high-level units. Salience methods such as layer-wise relevance propagation (ϵ-LRP),[38] DeepLIFT[39] and integrated gradients (IGs)[40] reflect the contributions of specific inputs to the final output, thus these methods must take the input values into account. DeepLIFT streamlines this computation by substituting the gradient at each non-linearity with its mean gradient in just one step. IG is a more comprehensive generalization of

Figure 3. Overall accuracy, class-specific accuracy and t-SNE visualization of the damaged MLP after the ablation of unit 12 in the first hidden layer. This unit is an example for the representation of features corresponding to many different classes, reprinted with permission from Meyes et al.[32]. MLP, multilayer perceptron; t-SNE. t-distributed Stochastic Neighbor Embedding.



DeepLIFT. It calculates the average partial derivative of each feature as the input transitions from a baseline to its ultimate value. SmoothGradient is another gradient-based method which bases the visualization on a smoothing of the gradient with a Gaussian kernel to mitigate the large fluctuation of local gradient.[41] Zhou et al proposed the class activation mapping (CAM), which is used to pinpoint distinguishing areas utilized by a specific class of image classification CNNs, which are void of any fully connected layers.[42] Selvaraju and team introduced a more versatile approach, Grad-CAM. It utilizes the gradient data channeled into the final convolutional layer of the CNN to assign significance to each neuron for a specific decision of interest.[43] These two are saliency methods but limited to CNN type network.

Although salience maps identify relevant regions for the importance of each pixel, it is a local explanation that is not generalizable to the whole data set and users has no control over what features being picked up by the method. For carrying out hypothesis testing on any user-understandable concept within a trained network and generating a global explanation, Kim and colleagues introduced the concept activation vectors (CAVs).[44] A vector in the activation space of a layer, representing a concept of human interest, is identified by comparing the activations in that layer generated by input examples from the concept set against random examples. SHAP values (SHapley Additive exPlanations), rooted in cooperative game theory, is a method employed to augment the transparency and interpretability of ML models.[45] SHAP's objective is to elucidate the prediction of an instance by determining the contribution of each feature to that prediction. SHAP has several variations, including Kernel SHAP and Tree SHAP, which is particularly suited for tree-based ML models such as decision trees, random forests, and gradient boosted trees.[46]

APPLICATION OF INTERPRETABILITY AI IN RADIATION ONCOLOGY

Image segmentation

Image segmentation is routinely performed, *e.g.* in radiotherapy, the delineation of organ at risks and target volume to be irradiated. The application of AI algorithms can potentially not only speed up such tedious activities, but also reduce differences in manual delineations arising from diverse expertise of the operators and the presence of broader contouring guidelines among institutions. It's important to note that most existing interpretability techniques are compatible with classification models, but they don't typically work with segmentation models. The most recent and remarkable effort was presented in 2021 by Chatterjee et al, who presented various interpretability techniques for segmentation models and used as a proof-of-concept experiments on a vessel segmentation model.[47] Their research specifically concentrated on input attributions and layer attribution methods, which highlight the crucial features of the image recognized by the model. Employing these interpretability techniques enables visualization of key features or regions in MRI images that the model deems vital in establishing the output, such as the Lenticulostriate arteries. Out of the 25 tested methods, they found out that attributions from the following methods had the most promising results: (A) guided backpropagation, (B) IGs, and (C) layer activation with guided backpropagation.

Computer-assisted diagnosis, prognosis, and predictions

AI algorithms for computer-assisted diagnosis, prognosis and prediction try to correlate the input with an outcome of interest (*e.g.* survival, histology grade). Linear regression, logistic regression are commonly used inherent interpretable models in different fields of healthcare like urology,[48,49] toxicology,[50] cardiology[51,52] or psychiatry.[53] Decision trees are very popular since the final model reproduces the logic of thinking of humans, with a set of pre-determined rules. These techniques have recently been employed for interpretive purposes in predicting

health-related conditions, which include occupational diseases[54] and knee osteoarthritis,[55] as well as gauging the severity of chronic diseases such as diabetes or Alzheimer's disease,[56] and mortality rates associated with conditions like myocardial infarction or perinatal stroke.[57]

However, for complex models like neural networks, post-training models should be used. For instance, SHAP was utilized to make sense of predictions regarding the prevention of hypoxemia during surgical procedures.[58] This application improved the anticipation of hypoxemia events by anesthesiologists by 15%. A recent application tested the introduction of domain knowledge by experts on feature subspaces.[59] The Model Understanding through Subspace Explanations (MUSE) was employed to elucidate decisions derived from a 3-level neural network trained on a depression diagnosis data set. This tool creates sets of if-then directives that globally describe the model's decisions. Additionally, it delivers a distinct collection of rules for a subspace, which are generated based on the attributes selected by a healthcare expert involved in patient care. In a recent study by Hosni et al, a DL network was used for mortality risk stratification based on standard-of-care CT images from non-small cell lung cancer (NSCLC) patients.[60] Utilizing Grad-CAM, it was noticed that the network was predominantly focusing on the juncture between the tumor and stroma (parenchyma or pleura). The most significant influences on predictions were presented as extensive continuous regions of relatively elevated CT density, extending into and beyond the tumor. Contrarily, regions with lower CT density were found to contribute minimally to the predictions. Cui et al adopted Grad-CAM methods to understand a DL model which were developed to predict radiation pneumonitis and local control for NSCLC patients.[61] Specifically, Grad-CAM was able to show that radiation dose around 20 Gy is crucial for predicting radiation pneumonitis, and inflammation-related cytokines contributed to the prediction of radiation pneumonitis (RP) and PET tumor radiomics play a role in determining the local control of radiation therapy (RT).

Some study also used DL networks as a feature extractor *per se* and then fall in the interpretability approaches presented for the ML algorithms. The most used approaches are encoder–decoder architectures, where only the encoding part of the network is used to extract features from the images. Some studies have shown that there exist correlations among "deep features" and human-defined features. An illustrative example is the research conducted by Zhang et al, where they employed CT images from pancreatic ductal adenocarcinoma (PDAC) patients gathered from two separate hospitals. Their findings revealed that the majority of transfer learning features bear a weak linear correlation with radiomics features, hinting at a potential synergistic relationship between these two sets of features.[62] Similar results have been obtained in low-dose CT images from the National Lung Screening Trial (NLST).[63] The approach used in this study involved the combination of features from pre-trained CNNs, CNNs trained on NLST data, and classical radiomics to construct classifiers. The highest level of accuracy, which stood at 76.79%, was achieved when these feature combinations were utilized. However, in some cases, the correlation between DL features and other features may indicate that CNN didn't learn any useful information. In a study focused on detecting pneumonia through chest radiographs, it was discovered that the CNN's predictions were not solely based on pathological features. Instead, they were more influenced by the metallic token placed on the patient's shoulder to indicate laterality, as revealed by the activation heat map method. Additionally, the research found that the metal token's location is site-specific and the hospital system's pneumonia prevalence significantly impacts the prediction accuracy for this task.[64]

Radiation therapy planning

Similar to AI for segmentations, the main focus of the studies in RT planning has been the plan automation, including beam directions and dose distributions. We didn't find studies that specifically focused on the investigation of the interpretability of the algorithms used to generate the plans. However, the literature is showing a growing interest towards a particular class of DL algorithms called "reinforcement learning" (RL). Although the aims or goal of RL is not exactly interpretability, but it does help with understanding the problem. In a study by Zhang et al, the investigators have formulated the interactions between the planner and the treatment planning system (TPS) into a finite-horizon reinforcement learning model.[65] Firstly, they assessed the planning status features, which were based on the expertise of human planners and defined as planning states. Next, they characterized planning actions to signify the common steps that planners would employ to address various planning needs. Lastly, in adherence to the RL methodology, they established a "reward" system reliant on an objective function steered by physician-imposed constraints. The team used visualization techniques to illustrate the learning enhancements of the system. For instance, the RL bot learned that if both the PTV33Gy coverage and stomach D1cc constraints are jeopardized, it's preferable to contemplate adding a lower constraint to a supporting structure, bypassing the intersection between the PTV and the stomach. Analyzing these planning strategies carries multiple advantages. It validates the reliability of the trained RL bot, as it yields consistent planning strategies on par with human planners. Additionally, the development of the action-value function provides meaningful insights into the learned planning strategies, fostering a knowledge map. Additionally, this method paves the way for examining "black box" models, enabling the discovery of innovative planning approaches that can enhance overall plan quality.

EVALUATION OF INTERPRETABILITY

Doshi-Velez and Kim proposed a taxonomy for the evaluation of interpretability.[4] The evaluation scheme was classified into three classes—application-grounded, human-grounded, and functionally grounded, based on evaluation complexity and risk associated to model application. Application-grounded evaluation (AGE) involves domain experts to evaluate interpretability in a real-life application, human-grounded evaluation (HGE) involves lay human evaluators, not necessarily domain experts in a simplified task that maintains only the essence of the target application, and functionally grounded evaluation (FGE) involves no humans but some form of definition of interpretability as a proxy. For

Table 1. SaMD categories

	Significance of information provided by SaMD to healthcare decision		
State of healthcare situation or condition	Treat or diagnose	Drive clinical management	Inform clinical management
Critical	IV	III	II
Serious	III	II	I
Non-serious	II	I	I

medical field, model's associated risk can be categorized into four classes, as listed in Table 1 [66], based of FDA guidelines that primarily considers two major factors: (a) significance of the information provided by the SaMD to the healthcare decision (*i.e.* treat or diagnose, drive clinical management, and inform clinical management) and (b) state of the healthcare situation or conditions (*i.e.* critical, serious, and non-serious). For example, SaMD designed to treat or diagnose patients in critical conditions are categorized into the Class IV while SaMD designed to inform clinical management for non-serious conditions are categorized into the Class I.[67,68]

AGE is the most complex as it involves human expert evaluators. In AGE, the objectives of a ML system are directly tested by the intended user group and is most appropriate for applications with higher associated risk such as clinical decision-making tools[69] and treatment planning tools.[70] In AGE of adaptive RT, doctors might evaluate the AI's dose adaptation decision. In general, AGE must have a high standard of experimental design, as its result may be crucial for a ML and DL system's approval of use. AGE may be applied for low-risk applications as well, however the high cost of AGE may not be justified. For such applications, HGE might be more appropriate where a non-expert human evaluator experiments with ML and DL system for simpler version of the targeted task. Doshi-Velez and Kim state that HGE becomes particularly attractive when conducting experiments with the target user group proves to be both difficult and costly. By involving lay-human HGE allows for a larger test size and lower expenses. HGE mainly assesses broader concepts of an explanation's quality, which may be adequate for applications with low risk. For minimal risk ML and DL system, FGE is the most appropriate form of evaluation. FGE is the least costly and is most appropriate once a class of ML and DL models are validated via HGEs. For instance, FGE may demonstrate the enhancement in prediction performance of a model that has already been established as interpretable.

Except Doshi-Velez and Kim's work, Brocki and colleagues examined the efficacy of various importance estimators in clarifying the classification of CT images by a convolutional deep network. They employed three unique evaluation metrics for this investigation.[71]

REGULATION OF AI TECHNOLOGY

The necessity for AI technologies to be regulated is widely recognized. Parikh et al laid out five criteria to inform the regulation of devices built on AI and predictive analytics, namely: significant end points, suitable benchmarks, interoperability and generalizability, specific interventions, and auditing mechanisms.[72] In the USA, the FDA introduced a regulatory framework for AI's application in medicine in April 2019 and the Digital Health Innovation Action Plan in January 2021.[73] This approach is directed towards software approval as a medical device, including AI technologies. The US policy avoids extensive, overarching regulation of AI, instead delegating specific responsibilities to individual federal agencies with the broader goal of preventing overregulation and fostering innovation.[74]

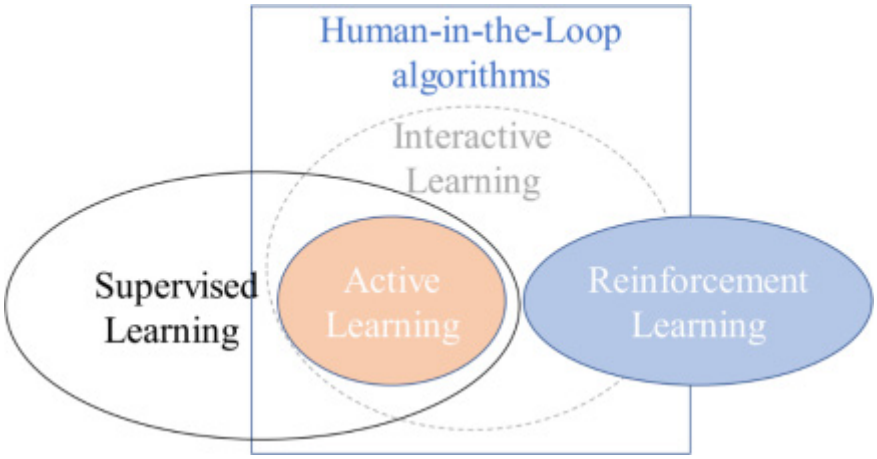
On the other hand, the European Union enforces regulations on algorithms making decisions based on user-specific indicators. Users hold the right to request a justification for a decision made about them by an algorithm, a privilege often termed as the "right to explanation".[75] The EU's regulatory approach for AI applications in medicine combines industry-specific and cross-sectional regulations. The European Commission recently put forward a proposal for a robust legal framework for AI, known as the AI Act, aiming to encourage AI adoption and cultivate an ecosystem of trust.[76] Applications of AI in medicine must meet the AI Act's requirements, as well as those specified under the current EU Medical Device Regulation.

Vokinger and colleagues drew parallels and distinctions between US and EU regulations for AI in medicine, focusing on three areas of transatlantic regulation: lifecycle regulation, algorithmic bias, and user transparency.[77] With future advancements in AI technology regulation in the medical field, interpretability will gain increased significance for complex systems.

DISCUSSION

Different interpretable methods have their advantages and disadvantages. There is no perfect method for all the models and tasks. Thus, it is important to understand their limitations and apply the optimal method for specific questions. The first step is always defining the question and understanding the data structure (explaining data). Depending on the question and available data, we can determine the models and then the best way to make the model interpretable. Though inherent interpretable models are easier to build and interpret, it is limited by the type of input data it can handle and the flexibility for complex interactions compared to DL methods. A linear model does not inherently possess interpretability. Just because a transparent model, like a decision tree, is built on manually crafted features does not make it interpretable by default, even if these individual features are grounded in specific domain knowledge and can be comprehended by a human. The model's interpretability is directly influenced by the quantity and complexity of its features. For instance, a linear model composed of thousands of parameters can be challenging to comprehend, thus limiting its interpretability. In comparison, DL is more flexible, which can handle varieties of inputs (images, videos). It can also model very complicated relations. Generally, if we have a small data set and small number of features, it makes sense to use inherent interpretable models. However, if we handle complex feature interactions, it

Figure 4. Foundations of human-in-the-loop algorithms, reprinted with permission from Otunaiya and Muhammad.[49]



is better to explore random forests, DL or other methods and do post-training interpretation.

For post-training methods, there are still challenges. While ablation test methods are relatively straightforward, due to their ability to directly measure the marginal effect of specific features on the output, and also because they are model-agnostic (meaning they can be applied to any black-box models), their application is not without limitations. These limitations are primarily associated with the selection of ablation techniques. For example, a simple question is how we choose the occluding box size when we occlude some regions of input images for a CNN model to interpret the results. The absence of a theoretically based method for selecting the ablation technique calls into question the dependability of the explanations produced as a result.[78] The time cost is another limitation for ablation methods.

Attribution-based methods have the advantage of being efficient and easy to implement, but they are also prone to noisy data points. IG improved the robustness but becomes much slower in computation. Another issue for the attribution method is the selection of baselines. Although zero baseline is often used and usually the output for zero input is near zero, this is still arbitrary selection. Sometimes, the expected value of each feature over the training set is used as a baseline.[45]

Beyond the scope of inherently interpretable models or post-training interpretation, data scientists have also proposed a hybrid interpretable model approach. This blends a complex black-box model with an inherently interpretable model to construct an interpretable model boasting performance comparable to the black-box model. For example, Gallego and his team showcased a clustering-based k-nearest neighbor classification that married an approximated similarity search method with the clustering model to reduce the complexity of the k-nearest neighbors algorithm.[79] Human-in-the-loop machine learning is another way to enhance interpretability of the built models. The human-in-the-loop machine learning is a hybrid methodology that merges data-driven and knowledge-driven approaches. This strategy incorporates *a priori* expert knowledge into ML frameworks as a means to mitigate model uncertainty.[80,81] In addition, it gives the right value to the knowledge producers and promotes interpretable learning machines and AI systems with a clear knowledge life cycle.[82] A brief summary of HITL algorithm is shown in Figure 4.

While numerous interpretation methods exist, certain general mistakes could lead to incorrect conclusions if these methods are improperly applied. Molnar et al identified some common pitfalls such as employing interpretation techniques in an inappropriate context, interpreting models with poor generalization, neglecting feature dependencies, interactions, uncertainty estimates, and issues arising in high-dimensional settings, or making unsupported causal interpretations. They provided examples to further illustrate these pitfalls.[83] In addition, Ennab et al noted certain limitations of current interpretability methods, such as the lack of model independence, inability to provide individualized feature importance, and failure to identify relevant feature sets for a single instance.[84] Being aware of these limitations and select or design the interpretable methods fit specific task is a necessary first step before we can leverage these advanced techniques to help improve patient healthcare safely and efficiently. Furthermore, to guarantee that medical AI fulfills its potential, it's crucial to raise awareness among developers, healthcare professionals, and legislators about the challenges and limitations of obscure algorithms in medical AI. This will necessitate fostering multidisciplinary collaboration as we progress.[85]

CONCLUSION

Interpretability of AI systems have become crucially important in medicine to provide safe support for clinical practice. We have summarized three main ways to help interpret AI models, discussed how to evaluate an interpretation method, and presented some applications in different fields of medicine. We also revealed some limitations of the current methods and suggest researchers to carefully understand the limitations of different methods and select appropriate methods for their specific tasks.

REFERENCES

1. Barragán-Montero A, Bibal A, Dastarac MH, Draguet C, Valdés G, Nguyen D, et al. Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model Interpretability, Explainability and data-model dependency. *Phys Med Biol* 2022; **67**(11). https://doi.org/10.1088/1361-6560/ac678a
2. Brouwer CL, Dinkla AM, Vandewinckele L, Crijns W, Claessens M, Verellen D, et al. Machine learning applications in radiation oncology: Current use and needs to support clinical implementation. *Phys Imaging Radiat Oncol* 2020; **16**: 144–48. https://doi.org/10.1016/j.phro.2020.11.002
3. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet Digit Health* 2021; **3**: e195–203. https://doi.org/10.1016/S2589-7500(20)30292-2
4. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. 2017: 170208608.
5. Papadimitroulas P, Brocki L, Christopher Chung N, Marchadour W, Vermet F, Gaubert L, et al. Artificial intelligence: deep learning in Oncological Radiomics and challenges of Interpretability and data harmonization. *Phys Med* 2021; **83**: 108–21. https://doi.org/10.1016/j.ejmp.2021.03.009
6. Iantsen A, Visvikis D, Hatt M. Squeeze-and-excitation normalization for automated delineation of head and neck primary tumors in combined PET and CT images. 3D Head and Neck Tumor Segmentation in PET/CT Challenge: Springer; 2021. p. 37-43.
7. Dercle L, Zhao B, Gönen M, Moskowitz CS, Firas A, Beylergil V, et al. Early Readout on overall survival of patients with Melanoma treated with Immunotherapy using a novel imaging analysis. *JAMA Oncol* 2022; **8**: 385–92. https://doi.org/10.1001/jamaoncol.2021.6818
8. Hader M, Frey B, Fietkau R, Hecht M, Gaipl US. Immune biological Rationales for the design of combined radio-and Immunotherapies. *Cancer Immunol Immunother* 2020; **69**: 293–306. https://doi.org/10.1007/s00262-019-02460-3
9. Flavell RR, Evans MJ, Villanueva-Meyer JE, Yom SS. Understanding response to Immunotherapy using standard of care and experimental imaging approaches. *Int J Radiat Oncol Biol Phys* 2020; **108**: 242–57. https://doi.org/10.1016/j.ijrobp.2020.06.025
10. Luo Y, Tseng H-H, Cui S, Wei L, Ten Haken RK, El Naqa I. Balancing accuracy and Interpretability of machine learning approaches for radiation treatment outcomes modeling. *BJR|Open* 2019; **1**: 20190021. https://doi.org/10.1259/bjro.20190021
11. 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. https://doi.org/10.1038/s42256-019-0048-x
12. Been K, Doshi-Velez F. Introduction to interpretable machine learning. Proceedings of the Tutorial on Interpretable Machine Learning for Computer Vision (Salt Lake City, Utah, United States, 2018), Conference on Computer Vision and Pattern Recognition; 2018.
13. Lipton ZC. The Mythos of model Interpretability: in machine learning, the concept of Interpretability is both important and slippery. *Queue* 2018; **16**: 31–57. https://doi.org/10.1145/3236386.3241340
14. Combi C, Amico B, Bellazzi R, Holzinger A, Moore JH, Zitnik M, et al. A manifesto on Explainability for artificial intelligence in medicine. *Artif Intell Med* 2022; **133**: 102423. https://doi.org/10.1016/j.artmed.2022.102423
15. Sammon JW. A Nonlinear mapping for data structure analysis. *IEEE Trans Comput* 1969; **C–18**: 401–9. https://doi.org/10.1109/T-C.1969.222678
16. Hinton GE, Roweis S. Stochastic neighbor Embedding. *Adv Neural Inf Process Syst* 2002; **15**.
17. Balasubramanian M, Schwartz EL. The Isomap algorithm and Topological stability. *Science* 2002; **295**: 7. https://doi.org/10.1126/science.295.5552.7a
18. Roweis ST, Saul LK. Nonlinear Dimensionality reduction by locally linear Embedding. *Science* 2000; **290**: 2323–26. https://doi.org/10.1126/science.290.5500.2323
19. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, et al. Dimensionality reduction for Visualizing single-cell data using UMAP. *Nat Biotechnol* 2019; **37**: 38–44. https://doi.org/10.1038/nbt.4314
20. Kobak D, Berens P. The art of using t-SNE for single-cell Transcriptomics. *Nat Commun* 2019; **10**: 5416. https://doi.org/10.1038/s41467-019-13056-x
21. Van der MaatenL, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008; **9**.
22. Abdelmoula WM, Balluff B, Englert S, Dijkstra J, Reinders MJT, Walch A, et al. Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data. *Proc Natl Acad Sci USA* 2016; **113**: 12244–49. https://doi.org/10.1073/pnas.1510227113
23. Hozumi Y, Wang R, Yin C, Wei G-W. UMAP-assisted K-means clustering of large-scale SARS-Cov-2 Mutation Datasets. *Comput Biol Med* 2021; **131**: 104264. https://doi.org/10.1016/j.compbiomed.2021.104264
24. Kobak D, Linderman GC. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat Biotechnol* 2021; **39**: 156–57. https://doi.org/10.1038/s41587-020-00809-z
25. Weisberg S. Applied Linear Regression. In: *Applied linear regression*. Hoboken, NJ, USA: John Wiley & Sons; 14 January 2005. https://doi.org/10.1002/0471704091
26. Kleinbaum DG, Dietz K, Gail M, Klein M, Klein M Logistic regression: Springer; 2002.
27. Myles AJ, Feudale RN, Liu Y, Woody NA, Brown SD. An introduction to decision tree modeling. *J Chemometrics* 2004; **18**: 275–85. https://doi.org/10.1002/cem.873
28. Friedman JH, Popescu BE. n.d.).(Predictive learning via rule ensembles. *Ann Appl Stat*; **2**: 916–54. https://doi.org/10.1214/07-AOAS148
29. Webb GI, Keogh E, Miikkulainen R. Naïve Bayes. *Encyclopedia of Machine Learning* 2010; **15**: 713–14.
30. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 2006; **15**: 651–74. https://doi.org/10.1198/106186006X133933
31. Ohri N, Duan F, Snyder BS, Wei B, Machtay M, Alavi A, et al. Pretreatment 18F-FDG PET Textural features in locally advanced non–small cell lung cancer: secondary analysis of ACRIN 6668/RTOG 0235. *J Nucl Med* 2016; **57**: 842–48. https://doi.org/10.2967/jnumed.115.166934
32. Meyes R, Lu de M, Puiseau CW, Meisen T. Ablation studies in artificial neural networks. 2019.
33. Hampel FR. The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 1974; **69**: 383–93. https://doi.org/10.1080/01621459.1974.10482962
34. Koh PW, Liang P Understanding black-box predictions via influence functions. International conference on machine learning: PMLR; 2017. p. 1885-94.
35. Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K-R. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning.

In: *Gradient-based attribution methods. Explainable AI: Interpreting, Explaining and Visualizing Deep*. Cham: Springer; 2019., pp. 169–91. https://doi.org/10.1007/978-3-030-28954-6
36. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. European conference on computer vision: Springer; 2014. p. 818-33.
37. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: The all convolutional net. 2014.
38. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On Pixel-wise explanations for non-linear Classifier decisions by layer-wise relevance propagation. *PLoS One* 2015; **10**(7): e0130140. https://doi.org/10.1371/journal.pone.0130140
39. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. International conference on machine learning: PMLR; 2017. p. 3145-53.
40. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. International conference on machine learning: PMLR; 2017. p. 3319-28.
41. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. Smoothgrad: removing noise by adding noise. 2017.
42. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Las Vegas, NV, USA. ; 2016. pp. 2921–29. https://doi.org/10.1109/CVPR.2016.319
43. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 IEEE International Conference on Computer Vision (ICCV); Venice. ; 2017. pp. 618–26. https://doi.org/10.1109/ICCV.2017.74
44. Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). International conference on machine learning: PMLR; 2018. p. 2668-77.
45. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017; **30**.
46. Lundberg SM, Erion GG, Lee S-I. Consistent individualized feature attribution for tree ensembles2018.
47. Chatterjee S, Das A, Mandal C, Mukhopadhyay B, Vipinraj M, Shukla A, et al. Torchesegeta: framework for Interpretability and Explainability of image-based deep learning models. *Applied Sciences* 2022; **12**: 1834. https://doi.org/10.3390/app12041834
48. Zhang Y, Yuan P, Ma D, Gao X, Wei C, Liu Z, et al. Efficacy and safety of Enucleation vs. resection of prostate for treatment of benign Prostatic hyperplasia: a meta-analysis of randomized controlled trials. *Prostate Cancer Prostatic Dis* 2019; **22**: 493–508. https://doi.org/10.1038/s41391-019-0135-4
49. Otunaiya KA, Muhammad G. Performance of Datamining techniques in the prediction of chronic kidney disease. *Csit* 2019; **7**: 48–53. https://doi.org/10.13189/csit.2019.070203
50. Zhang Q, Caudle WM, Pi J, Bhattacharya S, Andersen ME, Kaminski NE, et al. Embracing systems toxicology at single-cell resolution. *Curr Opin Toxicol* 2019; **16**: 49–57. https://doi.org/10.1016/j.cotox.2019.04.003
51. Joshi M, Melo DP, Ouyang D, Slomka PJ, Williams MC, Dey D. Current and future applications of artificial intelligence in cardiac CT. *Curr Cardiol Rep* 2023; **25**: 109–17. https://doi.org/10.1007/s11886-022-01837-8
52. Feeny AK, Chung MK, Madabhushi A, Attia ZI, Cikes M, Firouznia M, et al. Artificial intelligence and machine learning in arrhythmias and cardiac electrophysiology. *Circ Arrhythm Electrophysiol* 2020; **13**(8): e007952. https://doi.org/10.1161/CIRCEP.119.007952
53. Obeid S, Haddad C, Akel M, Fares K, Salameh P, Hallit S. Factors associated with the adults' attachment styles in Lebanon: the role of Alexithymia, depression, anxiety, stress, burnout, and emotional intelligence. Perspect Psychiatr Care 2019; **55**: 607–17. https://doi.org/10.1111/ppc.12379
54. Di Noia A, Martino A, Montanari P, Rizzi A. Supervised machine learning techniques and genetic optimization for occupational diseases risk prediction. *Soft Comput* 2020; **24**: 4393–4406. https://doi.org/10.1007/s00500-019-04200-2
55. Jamshidi A, Pelletier J-P, Martel-Pelletier J. Machine-learning-based patient-specific prediction models for knee osteoarthritis. *Nat Rev Rheumatol* 2019; **15**: 49–60. https://doi.org/10.1038/s41584-018-0130-5
56. Karun S, Raj A, Attigeri G. Comparative analysis of prediction algorithms for diabetes. Advances in Computer Communication and Computational Sciences: Proceedings of IC4S. Springer. Vol. Volume 1; 2019. pp. 177–87. https://doi.org/10.1007/978-981-13-0341-8
57. Prabhakararao E, Dandapat S. A Weighted SVM Based Approach for Automatic Detection of Posterior Myocardial Infarction Using VCG Signals. 2019 National Conference on Communications (NCC); Bangalore, India. ; 2019. pp. 1–6. https://doi.org/10.1109/NCC.2019.8732238
58. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of Hypoxaemia during surgery. *Nat Biomed Eng* 2018; **2**: 749–60. https://doi.org/10.1038/s41551-018-0304-0
59. Lakkaraju H, Kamar E, Caruana R, Leskovec J. Faithful and Customizable Explanations of Black Box Models. AIES '19; Honolulu HI USA. ; 27 January 2019. . 131–38. https://doi.org/10.1145/3306618.3314229
60. Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, et al. Deep learning for lung cancer prognostication: a retrospective multi-cohort Radiomics study. *PLoS Med* 2018; **15**(11): e1002711. https://doi.org/10.1371/journal.pmed.1002711
61. Cui S, Ten Haken RK, El Naqa I. Integrating Multiomics information in deep learning architectures for joint actuarial outcome prediction in non-small cell lung cancer patients after radiation therapy. *Int J Radiat Oncol Biol Phys* 2021; **110**: 893–904. https://doi.org/10.1016/j.ijrobp.2021.01.042
62. Zhang Y, Lobo-Mueller EM, Karanicolas P, Gallinger S, Haider MA, Khalvati F. Improving Prognostic performance in Resectable Pancreatic Ductal adenocarcinoma using Radiomics and deep learning features fusion in CT images. *Sci Rep* 2021; **11**: 1–11. https://doi.org/10.1038/s41598-021-80998-y
63. Paul R, Hawkins SH, Schabath MB, Gillies RJ, Hall LO, Goldgof DB. Predicting malignant nodules by fusing deep features with classical Radiomics features. *J Med Imaging (Bellingham)* 2018; **5**(1): 011021. https://doi.org/10.1117/1.JMI.5.1.011021
64. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK, et al. Variable generalization performance of a deep learning model to detect pneumonia in chest Radiographs: a cross-sectional study. *PLoS Med* 2018; **15**: e1002683. https://doi.org/10.1371/journal.pmed.1002683
65. Zhang J, Wang C, Sheng Y, Palta M, Czito B, Willett C, et al. An interpretable planning Bot for Pancreas stereotactic body radiation therapy. *Int J Radiat Oncol Biol Phys* 2021; **109**: 1076–85. https://doi.org/10.1016/j.ijrobp.2020.10.019
66. Group ISW. Software as a Medical Device (SaMD): Application of Quality Management System. International Medical Device Regulators Forum. 2021.

67. Group ISW Software as a medical device": possible framework for risk categorization and corresponding considerations. International Medical Device Regulators Forum; 2014.
68. Sun W, Niraula D, El Naqa I, Ten Haken RK, Dinov ID, Cuneo K, et al. Precision radiotherapy via information integration of expert human knowledge and AI recommendation to optimize clinical decision making. *Comput Methods Programs Biomed* 2022; **221**: 106927. https://doi.org/10.1016/j.cmpb.2022.106927
69. Niraula D, Sun W, Jin J, Dinov ID, Cuneo K, Jamaluddin J, et al. Arclids: A clinical decision support system for AI-assisted decision-making in response-adaptive radiotherapy. *Sci Rep* 2023; **13**: 5279. https://doi.org/10.1038/s41598-023-32032-6
70. McIntosh C, Conroy L, Tjong MC, Craig T, Bayley A, Catton C, et al. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. *Nat Med* 2021; **27**: 999–1005. https://doi.org/10.1038/s41591-021-01359-w
71. Brocki L, Marchadour W, Maison J, Badic B, Papadimitroulas P, Hatt M, et al. Evaluation of importance estimators in deep learning classifiers for Computed Tomography. In: *Explainable and Transparent AI and Multi-Agent Systems: 4th International Workshop, EXTRAAMAS 2022, Virtual Event*. Revised Selected Papers: Springer; 2022., pp. 3–18. https://doi.org/10.1007/978-3-031-15565-9
72. Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive Analytics in medicine. *Science* 2019; **363**: 810–12. https://doi.org/10.1126/science.aaw0029
73. Food AD. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. Food Drug Admin. . Silver Spring2021.
74. House W. Guidance for regulation of artificial intelligence applications. *Memo Heads Exec Dep Agencies* 2020.
75. Goodman B, Flaxman S. European Union regulations on Algorithmic decision-making and a "right to explanation *AIMag* 2017; **38**: 50–57. https://doi.org/10.1609/aimag.v38i3.2741
76. Commission E. Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. Office for Official Publications of the European Communities Luxembourg; 2021.
77. Vokinger KN, Gasser U. Regulating AI in medicine in the United States and Europe. *Nat Mach Intell* 2021; **3**: 738–39. https://doi.org/10.1038/s42256-021-00386-z
78. Ancona M, Ceolini E, Öztireli C, Gross M. Gradient-based attribution methods. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. 2019:169-91.
79. Gallego A-J, Calvo-Zaragoza J, Valero-Mas JJ, Rico-Juan JR. Clustering-based K-nearest neighbor classification for large-scale data with neural codes representation. *Pattern Recognition* 2018; **74**: 531–43. https://doi.org/10.1016/j.patcog.2017.09.038
80. Wu X, Xiao L, Sun Y, Zhang J, Ma T, He L. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* 2022; **135**: 364–81. https://doi.org/10.1016/j.future.2022.05.014
81. Luo Y, Cuneo KC, Lawrence TS, Matuszak MM, Dawson LA, Niraula D, et al. A human-in-the-loop based Bayesian network approach to improve imbalanced radiation outcomes prediction for hepatocellular cancer patients with stereotactic body radiotherapy. *Front Oncol* 2022; **12**. https://doi.org/10.3389/fonc.2022.1061024
82. Zanzotto FM. Human-in-the-loop artificial intelligence. *Jair* 2019; **64**: 243–52. https://doi.org/10.1613/jair.1.11345
83. Molnar C, König G, Herbinger J, Freiesleben T, Dandl S, Scholbeck CA, et al. General pitfalls of model-agnostic interpretation methods for machine learning models. In: *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML*. Vienna, Austria, Revised and Extended Papers: Springer; 18 July 2022., pp. 39–68. https://doi.org/10.1007/978-3-031-04083-2
84. Ennab M, Mcheick H. Designing an Interpretability-based model to explain the artificial intelligence Algorithms in Healthcare. *Diagnostics (Basel)* 2022; **12**(7): 1557. https://doi.org/10.3390/diagnostics12071557
85. The Precise4Q consortium, Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in Healthcare: a Multidisciplinary perspective. *BMC Med Inform Decis Mak* 2020; **20**: 1–9. https://doi.org/10.1186/s12911-020-01332-6