Circulation: Cardiovascular Imaging

ADVANCES IN CARDIOVASCULAR IMAGING

Explainable Artificial Intelligence and Cardiac Imaging: Toward More Interpretable Models

Ahmed Salih, PhD; Ilaria Boscolo Galazzo, PhD; Polyxeni Gkontra, PhD; Aaron Mark Lee, PhD; Karim Lekadir, PhD; Zahra Raisi-Estabragh, PhD; Steffen E. Petersen, DPhil

ABSTRACT: Artificial intelligence applications have shown success in different medical and health care domains, and cardiac imaging is no exception. However, some machine learning models, especially deep learning, are considered black box as they do not provide an explanation or rationale for model outcomes. Complexity and vagueness in these models necessitate a transition to explainable artificial intelligence (XAI) methods to ensure that model results are both transparent and understandable to end users. In cardiac imaging studies, there are a limited number of papers that use XAI methodologies. This article provides a comprehensive literature review of state-of-the-art works using XAI methods for cardiac imaging. Moreover, it provides simple and comprehensive guidelines on XAI. Finally, open issues and directions for XAI in cardiac imaging are discussed.

Key Words: artificial intelligence ■ cardiac imaging techniques ■ diagnostic imaging ■ machine learning

Cardiac imaging refers to noninvasive techniques that are designed to assess cardiac structure and function.[1] Several imaging modalities are now available and have taken a central position in guiding diagnosis and treatment of cardiovascular diseases. These modalities include, but are not limited to, cardiac magnetic resonance (CMR) imaging, cardiac computed tomography, and echocardiography.[2] The quality and speed of image acquisition in cardiac imaging has improved markedly over the preceding 2 decades. Advances in cardiac imaging have led to a huge influx of imaging data, necessitating the development of automatic tools for the efficient conversion of raw imaging data into meaningful clinical information. Large research cohorts, such as the UK Biobank, are increasingly incorporating cardiovascular imaging into their protocols.[3] Automated and scalable image analysis tools are thus essential for the extraction of meaningful phenotypes from imaging data banks.

Artificial intelligence (AI) has become a determining factor in this respect, showing extraordinary success and progress in a diverse range of domains. There are many areas within which AI might narrow the gap between clinical care and the abundance of cardiac imaging data by, for example, improving image acquisition, segmentation, feature extraction, diagnosis, and as a component of clinical decision support systems.[4] Machine learning (ML) is a branch of AI that does not require explicit programming.[5] ML methods are used widely in cardiac imaging studies to perform tasks including estimation of biological heart age,[6] heart disease risk prediction,[7] and heart failure (HF) prediction.[8]

Deep learning (DL) is a subset of ML that mimics the human brain to perform simple and complex tasks more efficiently than any other system. DL models do not require feature engineering, as is the case for classical ML models.[9] Deep neural networks, convolutional neural networks (CNNs), recurrent neural networks, and generative adversarial networks are the most common DL models. However, such models do not expose the rationale for their outcomes to human users. This absence of transparency hinders their adoption in clinical decision-making.[10] Indeed, in cardiovascular medicine, it is essential that explanations regarding AI results are offered in order for the users to understand, trust, and safely use their new partners.[11] This has led to the emergence of explainable AI (XAI). The goal of XAI is to enhance the interpretation

Correspondence to: Ahmed Salih, PhD, Queen Mary University of London, Leicester LE17RU, United Kingdom. Email a.salih@qmul.ac.uk
Supplemental Material is available at https://www.ahajournals.org/doi/suppl/10.1161/CIRCIMAGING.122.014519.
For Sources of Funding and Disclosures, see page 354.

Circulation: Cardiovascular Imaging is available at www.ahajournals.org/journal/circimaging

Nonstandard Abbreviations and Acronyms

AI	artificial intelligence
CAD	coronary artery disease
CMR	cardiac magnetic resonance
CNN	convolutional neural network
DL	deep learning
Grad-CAM	gradient-weighted class activation mapping
HF	heart failure
LV	left ventricle
MI	myocardial infarction
ML	machine learning
RV	right ventricle
SHAP	Shapley additive explanations
XAI	explainable artificial intelligence

of black box model outputs, while maintaining high levels of performance.[12] Therefore, XAI is essential to ensure core aspects of trustworthiness.[10] Here, we will provide a concise review of the concepts of XAI for clinicians, without AI backgrounds, and will detail the usages and advantages of XAI in cardiac imaging. Finally, we will discuss open issues and directions for future research.

XAI LANDSCAPE

Recent advances have led to an explosion of AI applications in many fields. However, improvements in model performance have come at the cost of increased technical complexity. While AI models are believed to have the potential to revolutionize health care delivery, real-world deployment and adoption requires the trust of practitioners, patients, and stakeholders. In the last 10 years, there has been an enormous surge in the use of the term XAI, mirroring advances in AI, as illustrated in Figure 1. XAI is a set of tools and processes that enable human users to comprehend both how AI models work and how they reach a specific outcome.

Classic ML models (eg, linear regression, random forest, and decision tree) have reduced complexity compared with DL models and, as such, are considered more interpretable. However, DL-based models have been shown, generally, to outperform classic ML on the same tasks.[13] Difficulties in the interpretation of DL model outcomes have led to a lack of trust and represent a major obstacle in their translation to clinical use. Figure 2 illustrates the relationship between explainability and performance for the most widely used ML and DL models. The ideal outcome is that the model to be deployed will have both high performance and high interpretability, but this can be rarely accomplished in real-world scenarios, requiring a trade-off between model performance and interpretability.

The terms interpretable and explainable are often used interchangeably. However, Gunning et al[12] argue that their definitions are domain and task dependent. Broadly speaking, interpretability is the degree to which a human could understand or intuit how a model has reached a given outcome.[13] Explainability, on the other hand, relates to the internal mechanism and the logic of the ML system. Related and commonly coencountered terms are as follows: trustworthiness, the degree of confidence that the model would act as intended when encountering a given task; and transferability, the degree to which model outcomes can be adapted to a different domain. While there is no standard agreement on what an explainable model should comprise, research activities in XAI focus on ensuring trustworthiness, transferability, fairness, accessibility, and interactivity.[10]

TAXONOMY OF XAI

Interpretable models can be classified into several categories based on different criteria.[14] The first criterion is whether the interpretability is intrinsic (ante hoc) or established after training (post hoc). ML/DL models are considered intrinsic if they are self-explanatory and provide an immediate explanation on how the model has transformed the inputs to the output.[15] Decision trees and logistic regression are examples of simple intrinsic models as they are designed to provide not only classification/prediction results but also feature importance scores guiding their decision. In contrast, post hoc interpretability requires a second model to explain the current model after training. For instance, gradient-weighted class activation mapping (Grad-CAM)[16] is a post hoc XAI method that is usually applied to DL models. It visualizes attention map by generating a heat map on the nearest convolutional layer to the fully connected layers, as the last convolutional layer carries the spatial information that the model uses to make the decision. Saliency maps were proposed as attention visualization method by calculating the gradient of the output category with respect to input.

XAI models can be further classified into specific and agnostic. Model-specific refers to any explainable method that has been designed and developed to work with a single model or specific groups of algorithms. Usually, a model-specific method depends on the working and internal architecture (eg, the weights of a linear model) of the specified ML/DL models. On the other hand, model agnostic refers to explainable and flexible methods that can be applied to any model or algorithm,[17] irrespective of complexity.[13] Shapley additive explanations (SHAP) is an example of a model-agnostic method.

XAI methods provide either local or global explanations: local explanations hold for a specific or random instance, that is, a specific patient, whereas global explanations apply to all instances of the model.[13] Many

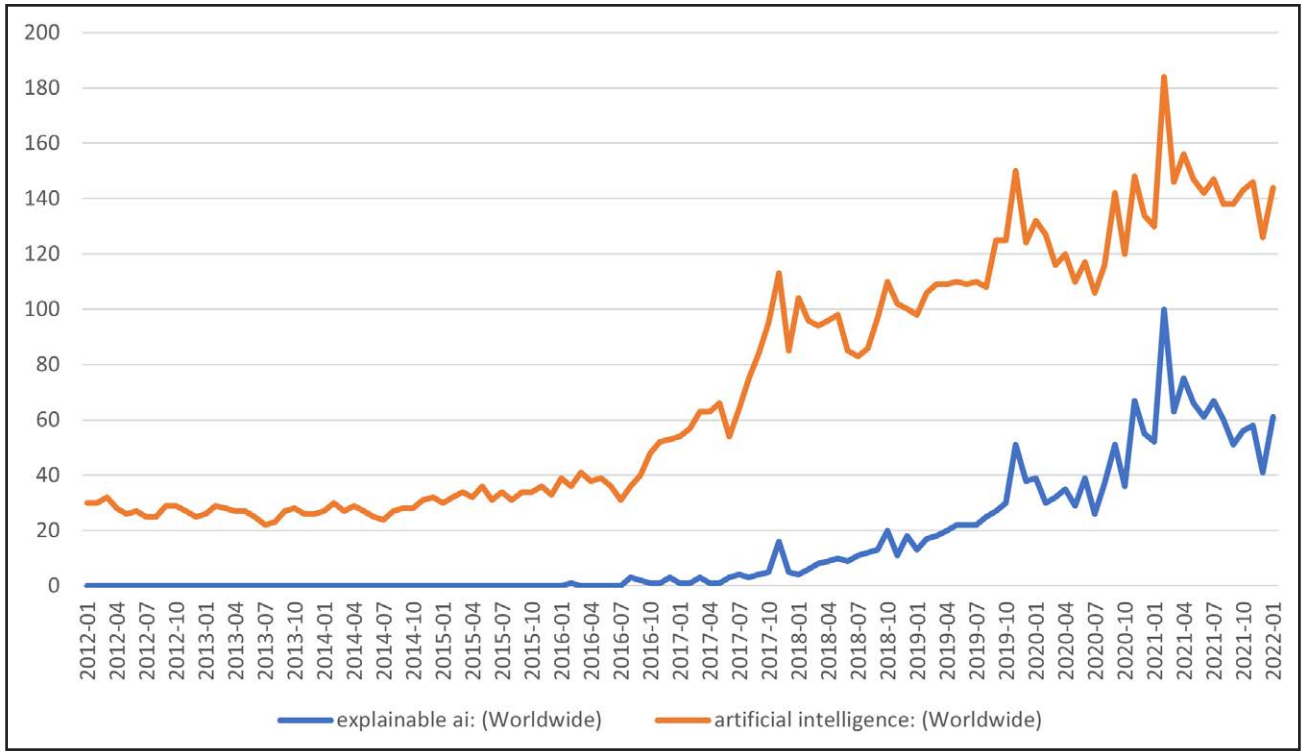

Figure 1. Google Trends is a free tool that analyzes the popularity of Google search terms using real-time data.
Google Trends shows the trend in search terms for artificial intelligence (AI)/explainable AI for the past 10 years. The *y* axis represents the normalized relative number of searches of the terms over time.

explainability methods are either built for a specific model, focus on specific data, or provide either local or global explanations: as such, there is no best-in-class method. It may be argued that SHAP is comprehensive in the sense that it is model agnostic and provides both local and global explanations.[13]

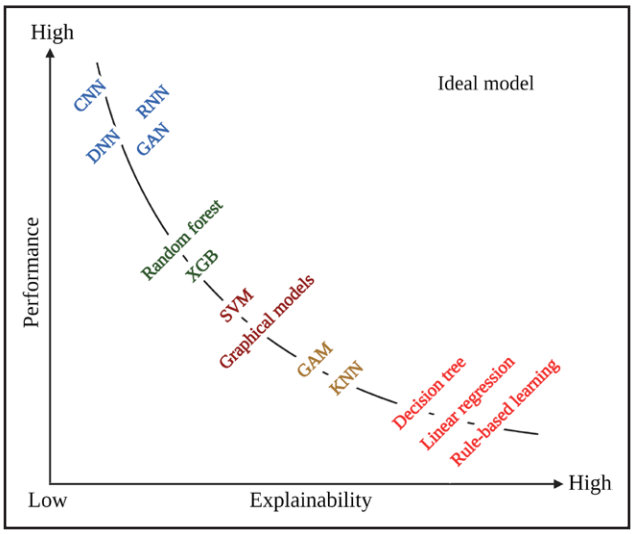

Figure 2. Model explainability versus performance for some selected artificial intelligence methods.
Graphical models: probabilistic model such as Bayesian network. Rule-based learning: any model uses rules (eg, if:then) to make a decision. CNN indicates convolutional neural network; DNN, deep neural network; GAM, generalized additive model; GAN, generative adversarial network; KNN, K-nearest neighbor; RNN, recurrent neural network; SVM, support vector machine; and XGB, extreme gradient boosting.

Figure 3A illustrates the outcome of 2 XAI methods (Grad-CAM and saliency map) that represent how the model can be distracted by motion artifact. The model was intended to segment myocardium; however, the attention map shows that the model was driven outside the myocardial region due to motion artifact. For instance, in segment 2 (a, b, d, and e), the attention map did not fully cover the inferoseptal and anteroseptal segment that could be related to the fact that the motion artifacts in these regions often occurred together. The red arrow represents the distraction. Figure 3B and 3C is an illustrative example of XAI model-agnostic methods that show the local and global explanation for a binary classification model to distinguish patients with myocardial infarction (MI) from the control group (non-MI) created from the UK Biobank. Figure 3A is the global explanation from SHAP, which illustrates the contribution of each predictor to the model outcome for all instances. It shows that high cholesterol, sex (male, 1), and body mass index are the most significant and contribute positively to the model outcome, indicating that small increases (change from female to male) to these values will increase the possibility of having MI. On the contrary, height contributes negatively. Figure 3B is a local explanation for a specific subject using the local interpretable model-agnostic

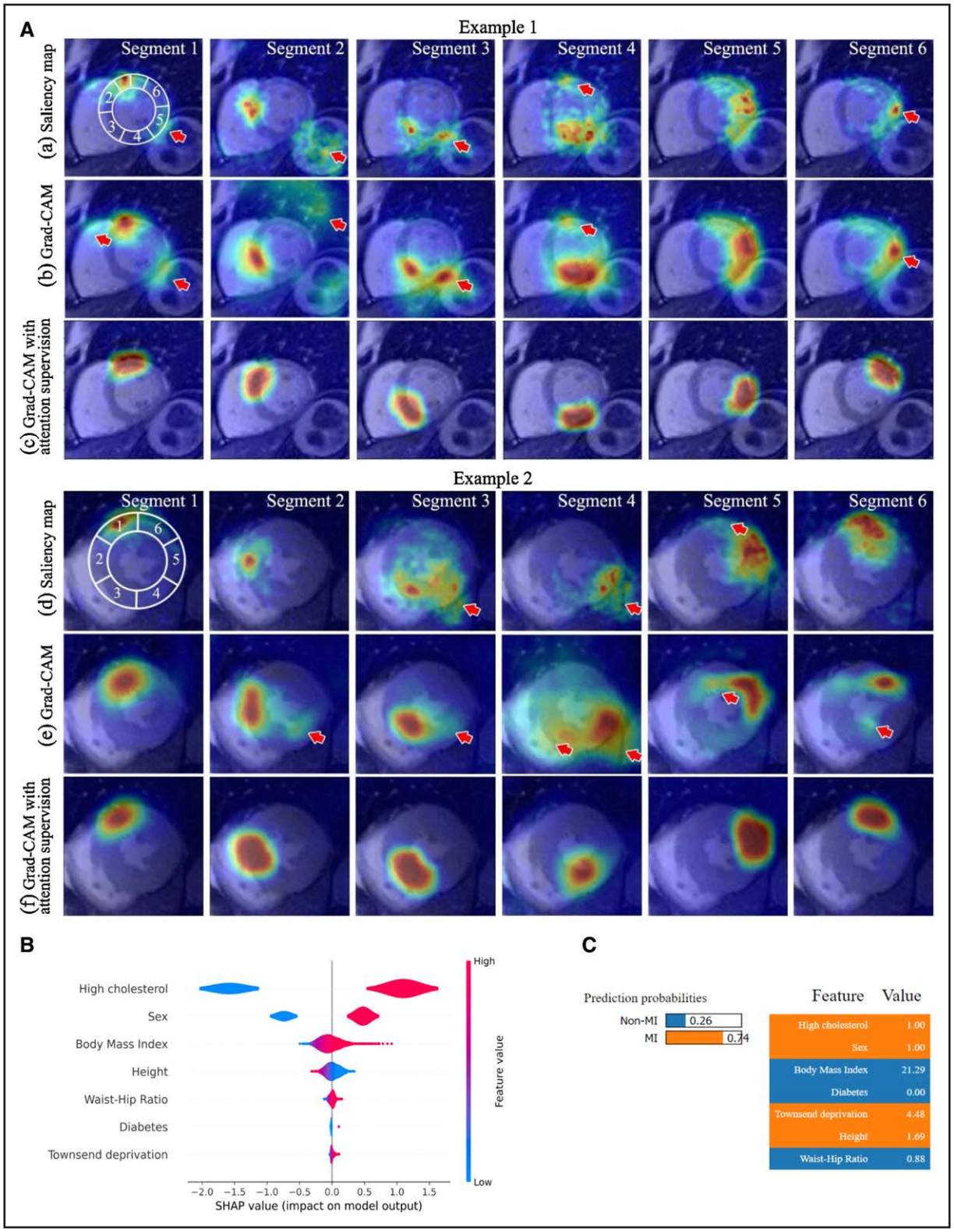

Figure 3. Clinical examples and interpretation.
A, Attention map generated by gradient-weighted class activation mapping (Grad-CAM) and saliency in cardiac T1 mapping. **B**, Global list of informative predictors from Shapley additive explanations (SHAP). High cholesterol, sex, and body mass index are contributing positively to model output (myocardial infarction [MI]) while height contributes negatively (non-MI). **C**, Local features contributions for a specific subject in the model. It shows the prediction probability for each class the subject might belong to. The color indicates whether the feature contributes to MI or non-MI classes while the numbers in the table represent the effect size in the model. **A**, Adapted from Zhang et al[18] with permission. Copyright © Elsevier, 2020.

explanations method. It shows the probability of the subject being part of the control group (0.26) or as having had an MI (0.74). In addition, it shows the contribution of each predictor for each class. For instance, having high cholesterol and being male contributes to the MI classification while having a body mass index value equal to 21.29 and the absence of diabetes contribute to the non-MI classification. Thus, methods such as SHAP and local interpretable model-agnostic explanations can be used to examine the under-the-hood workings of AI models and examine their biological plausibility. These results can be used to foster confidence in model outputs to end users. Other real examples on cardiac data are reported in Table S1.

XAI IN CARDIAC APPLICATIONS

Figure 4 illustrates the 2 common paths that cardiac imaging applications generally take, with the overall goal of leading to accurate diagnosis and prognostication. Of note, CMR was taken as an example over other modalities in this figure and in the subsequent description, although these concepts can be similarly extended to other cardiac imaging modalities. In image-based models, raw or minimally preprocessed images (2-dimensional/3-dimensional/4-dimensional depending on the model) are used directly as input, avoiding feature handcrafting. Usually, these are complex models (eg, CNNs and recurrent neural networks), and including an XAI method is required to make the model interpretable. The most common XAI methods generally applied to such data are SmoothGrad,[19] class activation mapping,[20] Grad-CAM,[16] and saliency maps.[21] The output of these XAI methods is usually a heat map, applied to a random or specific cardiac image from test data set, to represent the regions (pixels or voxels) that drive the model output.

On the other hand, feature-based methods require (1) preprocessing to extract the regions of interest from the images (for example, the left ventricles [LVs] and right ventricles [RVs]), a process commonly known as segmentation, and then (2) the derivation of measures representing local characteristics. Extracted features could be either conventional CMR or radiomics. Conventional CMR generally extracts LV/RV volumes in end diastole and end systole, LV/RV stroke volumes, LV/RV ejection fractions, and the LV mass.[22] Such approaches provide detailed information related to the size of cardiac chambers that can be used to distinguish healthy and pathological states.

Radiomics is a novel image analysis method that transforms raw cardiac images into quantifiable data representing the structural and functional characteristics of cardiac tissue.[23] It includes shape features such as volume, area, and surface; first-order features such as energy and entropy; and textural features such as contrast and spherical disproportion. Radiomics can be extracted in end diastole and end systole. A full description of all radiomics that can be extracted from cardiac images and the relevant open-source packages (Pyradiomics) can be found in the study by Van Griethuysen et al.[24] These handcrafted features are then used to train the model for a given application, using either regression, classification, or unsupervised learning tasks such as clustering. Subsequently, XAI methods can be applied to aid interpretation of the model output. The outcomes of XAI are usually a list of informative predictors or class membership probabilities (eg, HF versus control) and the contribution of each predictor for each class in a classification model. The Table 1 lists the most common XAI methods applied to both image-based and features-based models, together with a brief description related to their aims, properties, applicable input and ML/DL model, and applications in previous cardiac imaging studies.

LITERATURE REVIEW

We performed a literature search on December 9, 2022, within PubMed, Scopus, IEEE Xplore, and Web of Science databases, to determine the main papers available in the current literature on cardiac imaging and XAI. The query consisted of 4 main parts (Table S2) that combined different terminologies for cardiac (cardiac and heart), imaging methods (eg, CMR, computed tomography, and echocardiograms), ML/DL (eg, ML and CNN), and for a wide range of XAI models (eg, SHAP, Layer-wise Relevance Propagation, Grad-CAM, and Local Interpretable Model-agnostic Explanations). The research was restricted to title and abstract, and we considered only articles written in English, with no time restriction. This led to a total of 49 unique articles. Of these, 17 articles were removed after full text review as they did not specifically include XAI methods and the words explainable/interpretable were only used in a more general sense. Therefore, 32 original articles (24 journal papers and 8 international conference proceedings) were considered, suggesting that the application and development of XAI methods in cardiac imaging is still in its infancy (Table 1; Table S3).

Sixteen of these articles relied on well-known XAI methods and applied them to interpret the outcomes of complex DL-based models (more scarcely ML based) having cardiac images from CMR, nuclear medicine, echocardiography, or histopathology as inputs. The aims of these articles were diverse and included the automatic calculation of the LV volume and ejection fraction[28,35]; the automatic segmentation of cardiac structures as LV/RV, myocardium, interventricular septal, and posterior LV wall[39,51,55]; and classification approaches. Classification was investigated in the context of cardiac disorders such as mitral valve diseases,[48] myocardial injury,[50] coronary artery disease (CAD),[34,54] cardiomyopathy,[33] congenital

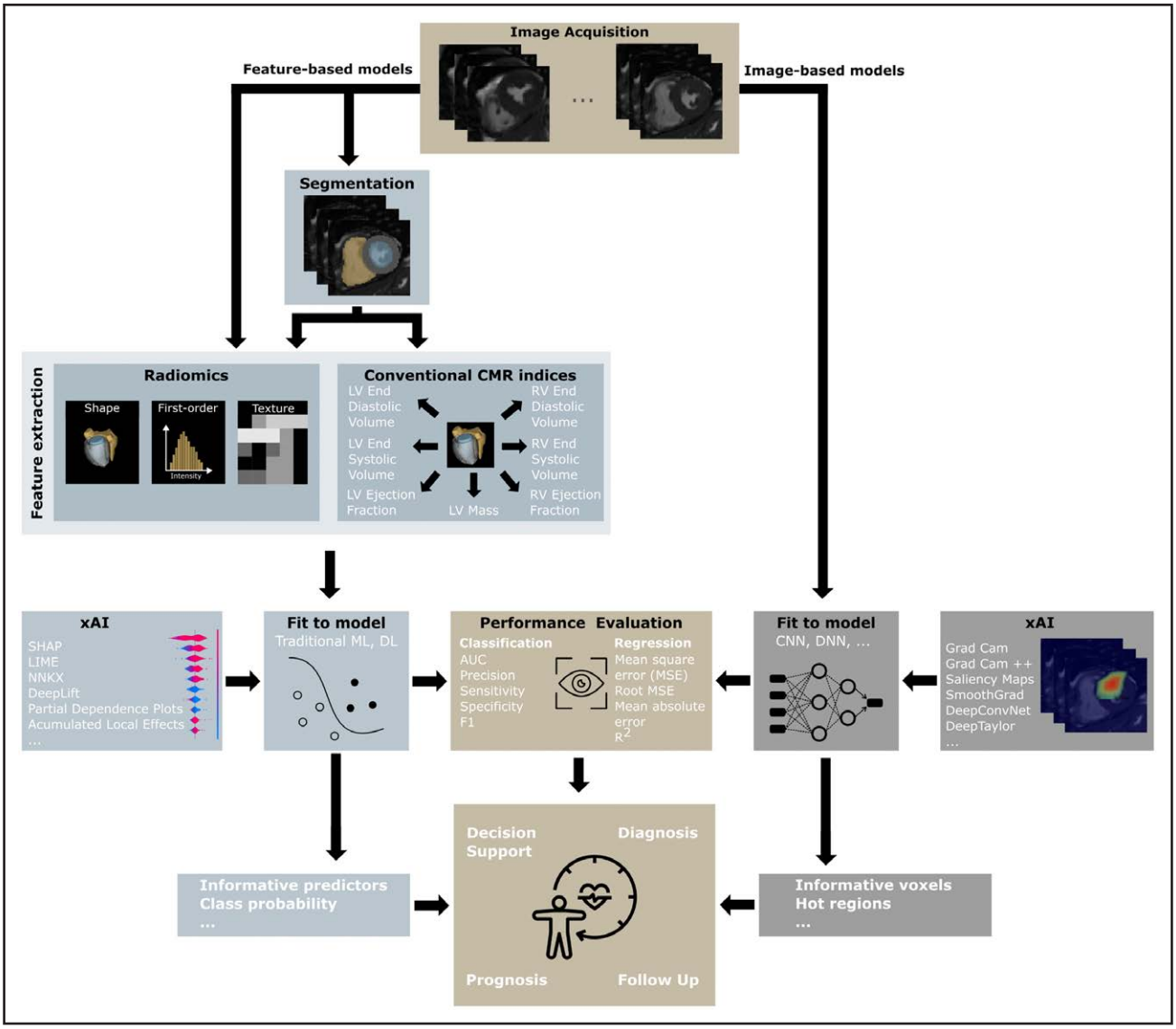

Figure 4. Overview of the 2 common paths that cardiac imaging studies might follow using cardiac magnetic resonance (CMR) as an example.

AUC indicates area under the curve; CNN, convolutional neural network; DeepConvNet, deep learning with convolutional neural networks; DL, deep learning; DNN, deep neural network; Grad-CAM, gradient-weighted class activation mapping; LIME, local interpretable model-agnostic explanations; LV, left ventricle; ML, machine learning; MSE, mean squared error; NNKX, neural network knowledge extraction; RV, right ventricle; SHAP, Shapley additive explanations; and XAI, explainable artificial intelligence. Reproduced with permission. Copyright © 2023, the UK Biobank.

heart disease,[40] and HF[32]; even in newborn[49] for the presence of intracardiac devices (eg, catheters, pacemaker, and defibrillator leads) or motion[18]; for severe left atrial dilation and LV hypertrophy[43]; and for CMR image view.[27] In order to improve the interpretability of the segmentation/prediction/classification results, these articles generally focused on DL-based XAI methods as class activation mapping or Grad-CAM, SmoothGrad, saliency maps, testing with concept activation vectors, and guided backpropagation. Class activation mapping–based methods are often used when 2- or 3-dimensional images are available, and some extensions have been proposed in cardiac imaging applications. For instance, Pérez-Pelegrió et al[28] developed a new explainable approach that combines class activation mapping with U-net to automatically estimate the LV volume in end diastole and obtain the result in the form of a segmentation mask without segmentation labels to train the algorithm.

Grad-CAM was used in 7 cardiac imaging studies, either for classification[18,34,40,48–50] or segmentation.[51] The latter in particular proposed a new interpretable CNN model (fast and accurate echocardiographic automatic segmentation based on U-Net) that integrates U-net architecture and transfer learning (from Visual Geometry Group 19) to segment 2-dimensional echocardiography of 88 patients into 3 regions (LV, interventricular septal, and posterior LV wall). The model showed better performance compared with traditional methods by achieving high mean dice score

Table 1. Taxonomy of Interpretability Methods and Related Properties

XAI	Aim	Properties	Input	Model	Application
PDP[25]	Shows the marginal effect of 1 or 2 predictors on the outcome	G, A	*	M, D	None
ALE[26]	Shows the average effect of features on the outcome	G, A	*	M, D	None
Class activation map[20]	Build discriminative image regions to show the regions used by the model	L, S	†	D	Classification[27]; Regression[28]
RxREN[29]	Extract rules that drive the model using classified and misclassified data	G, S	*	D	None
NNKX[30]	Knowledge extraction from multilayers trained model	L, S	*	D	None
SHAP[31]	Provide feature importance list locally and globally based on game theory	G, L, A	*†	M, D	Classification[32–34]; Regression[35]
LIME[36]	Explain the contribution of each feature toward the outcome for one single instance	L, A	*†	M, D	None
Layer-wise relevance propagation[37]	Generate a heat map in the input space to reveal the contribution of each voxel in the model outcome	L, S	†	D	None
Guided backpropagation[38]	Visualize the learning of the intermediate layer of deep learning models	L, S	†	D	Segmentation[39]; Classification[40]
DeepLIFT[41]	Shows the additive features attribution to the model outcome	L, S	*†	D	None
Seq2Seq[42]	Visualize and debug sequence-to-sequence tool	L, S	*†	D	None
SmoothGrad[19]	Improve the sensitivity maps generated on the input image by removing the noise	L, S	†	D	Classification and Regression[43]
Saliency maps[21]	Generate saliency maps, which shows the contribution of each pixel toward the model output	L, S	†	D	Classification[18]
DeepTaylor[44]	Generate heat maps, which shows the contribution of each pixel toward the model output	L, S	†	D	None
DeConvNet[45]	Generate heat maps, which shows the contribution of each pixel toward the model output	L, S	†	D	None
Pattern attribution[46]	Generate heat maps, which shows the contribution of each pixel toward the model output	L, S	†	D	None
Integrated gradients[47]	Generate heat maps, which shows the contribution of each pixel toward the model output	L, S	†	D	None
Grad-CAM[16]	Generate heat map, which shows the contribution of each pixel toward the model output	L, S	†	D	Classification[18,34,40,48,49,50]; Segmentation[51]
Grad-CAM++[52]	Improved version of Grad-CAM	L, S	†	D	None
TCAV[53]	Features attribution	L, S	†	D	Classification[54]; Segmentation[55]

Please refer to Table S3 for more details on the cardiac imaging papers. A indicates agnostic; ALE, accumulated local effects; D, deep learning; DeConvNet, deconvolution network; DeepLIFT, deep learning important features; G, global; Grad-CAM, gradient-weighted class activation mapping; L, local; LIME, local interpretable model-agnostic explanations; M, machine learning; NNKX, neural network knowledge extraction; PDP, partial dependence plot; RxREN, rule extraction by reverse engineering; S, specific; Seq2Seq, sequence-to-sequence models; SHAP, Shapley additive explanations; TCAV, testing with concept activation vectors; and XAI, explainable artificial intelligence.

*Tabular.

†Images.

coefficients (0.932, 0.848, and 0.868 for the 3 regions, respectively). Adding Grad-CAM to the model allowed the authors to visualize the feature maps from the U-net for each of the 3 regions and thus to highlight the areas contributing more to the segmentation outcome.

Testing with concept activation vectors[53] is an XAI method that was designed by Google AI researchers to address the issue of fairness and bias in AI models, focusing on high-level concepts (stripes in a zebra) rather than on low-level concepts (edges and lines in a picture). Testing with concept activation vectors was used twice in these cardiac imaging studies to interpret DL models. In a classification model based on variational autoencoder, it was used to link some concepts (eg, low LV ejection fraction) to the output (MI and CAD versus control).[54] Thereafter, they inspected the defined concepts and established whether they were significantly related to the model output by changing the concepts and observing the alterations in the image domain. Their results indicate that filling rates and LV ejection fraction had high contributions to the model output. In addition, testing with concept activation vectors was used to discover different

concepts for different cardiac conditions in a trained model to segment cardiac MRI into LV, RV, and myocardium in 5 subject groups including controls, chronic MI, hypertrophic cardiomyopathy, dilated cardiomyopathy, and dilated RV.[55] Their model detected a wide range of significant concepts to the model output; however, these concepts were similar in the 5 groups and require further clinical validation.

In addition, 4 studies relied on SHAP to interpret the model outputs.[32–35] One of these studies[32] relied on SHAP to develop and test an explainable ML model to assess whether noncontrast CMR could bring added value to predicting HF hospitalization compared with clinical data only. Their results demonstrated that CMR-based ML models could provide a significantly superior prediction of HF hospitalization (area under the curve, 0.81) compared with the basic clinical model (area under the curve, 0.64). Moreover, their SHAP analysis revealed left and right atrium strains as top imaging markers contributing to the more accurate prediction of future HF hospitalization.[32]

SHAP and Grad-CAM were also integrated in a single study to build an XAI model using myocardial positron emission tomography images to predict mortality risk.[34] Grad-CAM was used to create attention maps in the model, whereas SHAP was incorporated to determine the importance of input.

The remaining 16 articles proposed different approaches with the common aim of making their models more interpretable and to cast light on the complex network structure of the DL methods that impact the interpretation of the model outcomes. In particular, Guo et al[56] developed a new way to enhance cardiac magnetic resonance imaging multiclass segmentation by combining the strengths of CNN and interpretable ML algorithms, by generating labeling probability maps for each structure to guide the segmentation. Their results demonstrated the possibility of applying CNN even in small training data sets and at the same time improving the accuracy for cardiac MRI segmentation in both research and clinical patient care. Another article assessed the quality of echocardiograms in apical 4-chamber view and visualized the learnt features in the model using a novel technique.[57] They first assessed the quality of the echocardiograms by an expert cardiologist. Then, using the scores assigned by the experts, they trained a CNN to calculate an automatic echo score in real time. Overall, the embedding in their model of an interpretation method helped to map the learnt features from the convolutional layers to the images. A CAD-DL model was proposed by Otaki et al[58] and applied to single-photon emission computed tomography myocardial perfusion imaging to detect obstructive CAD, while spatial decomposition network[59] was proposed to factorize 2-dimensional images into spatial anatomical factors and nonspatial anatomical factors for segmentation in order to increase interpretability. Finally, another interpretable model was proposed to derive a score to represent the quality of the CMR reconstruction.[60] The authors proposed a semantic interpretability score that measured the visibility of the regions of interest in both ground truth and the reconstructed images. The score was calculated using mean dice overlap between the experts-segmented cardiac MRI (ground truth) and the segmented cardiac MRI by pretrained model (reconstructed images). Their method considers human perception into account when evaluating the quality of image reconstruction.

The complete list of the 32 articles reviewed in the current work with more details on data, study cohort, and application is given in Table S3.

OPEN ISSUES AND RESEARCH DIRECTION

Our literature review suggests limited use of XAI methods in cardiac imaging, although the last year has witnessed an increased number of articles on this topic. XAI is indispensable in applications where there is no visual feedback such as classification and regression while its role might become less important when the output of the model can be visually assessed such as in segmentation tasks. A list of open-source XAI tools currently available and that can be used in this domain is reported in Table S4. It is a desirable characteristic, particularly in health care applications, as it can combat the uncertainty caused by the black box nature of underlying algorithms and subsequent lack of trust. For example, XAI has been used to avoid unnecessary financial costs related to inappropriate prescribing[61] or long-term health care for those with stable CAD.[62] Furthermore, using XAI may help compliance with the implementation of general data protection, when AI is used in automated decision-making.[63]

Importantly, Ghassemi et al[64] have considered the relevance of XAI for health care applications in a broader context and argue that adoptability of medical AI should be driven by rigorous internal and external validation as, ultimately, this is necessary for deployment in a medical context. The authors conclude that XAI has more of a role at a developmental level through model troubleshooting and audit where it could be used to assess for bias, failure modes, or performance issues.

Notwithstanding an increasing uptake of XAI, there is no standard way to evaluate the outcomes that result from its use. Three levels of evaluation have been proposed to evaluate interpretability methods: (1) application-grounded, (2) human-grounded, and (3) functionally grounded evaluation.[65] Application-grounded evaluation uses experts to assess explainability but is expensive, nongeneralizable, and subjective. Human-grounded evaluation uses nonclinical evaluators and so is less expensive but is not applicable in medical fields due to the

privacy issues. Functionally grounded evaluation aims to evaluate XAI methods using proxies and statistical methods to quantify model uncertainty. Selectivity[37,66] and sensitivity[47,67] have been put forward as proxies to evaluate XAI methods: selectivity indicates how the model performance would be affected if significant features/regions identified by XAI method were removed; and sensitivity is related to feature perturbation and its effect on XAI outcomes. Our conclusion is that the evaluation of XAI methods is an emerging and young field of research and that, therefore, involving experts in the evaluation of XAI outcomes remains of crucial importance.

ARTICLE INFORMATION

Affiliations

William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University of London, United Kingdom (A.S., A.M.L., Z.R.-E., S.E.P.). Department of Computer Science, University of Verona, Italy (I.B.G.). Department of de Matemàtiques i Informàtica, University of Barcelona, Spain (P.G., K.L.). Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS Trust, London, United Kingdom (Z.R.-E., S.E.P.). Health Data Research UK, London (S.E.P.). Alan Turing Institute, London, United Kingdom (S.E.P.).

Sources of Funding

Dr Salih is supported by a British Heart Foundation project grant (PG/21/10619). Dr Boscolo Galazzo acknowledges support from Fondazione CariVerona (EDIPO project, reference No. 2018.0855.2019). Dr Raisi-Estabragh recognizes the National Institute for Health Research (NIHR) Integrated Academic Training programme, which supports her Academic Clinical Lectureship post and was also supported by British Heart Foundation Clinical Research Training Fellowship number FS/17/81/33318. S.E. Petersen and Dr Lee acknowledge support from the SmartHeart EPSRC programme grant (www.nihr.ac.uk; EP/P001009/1) and from the London Medical Imaging and Artificial Intelligence Centre for Value Based Healthcare (AI4VBH), which is funded from the Data to Early Diagnosis and Precision Medicine strand of the government's Industrial Strategy Challenge Fund, managed and delivered by Innovate UK on behalf of UK Research and Innovation. S.E. Petersen acknowledges support from the NIHR Biomedical Research Centre at Barts. S.E. Petersen, Dr Lekadir, and Dr Gkontra have received funding from the European Union Horizon 2020 research and innovation programme under grant agreement number 825903 (euCanSHare project). Dr Lekadir and Dr Gkontra have additionally received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101057849 (DataTools4Heart project).

Disclosures

S.E. Petersen provides consultancy to Cardiovascular Imaging, Inc, Calgary, Alberta, Canada. The other authors report no conflicts.

Supplemental Material

Tables S1–S4
References 68–78

REFERENCES

1. Blankstein R. Introduction to noninvasive cardiac imaging. *Circulation*. 2012;125:e267–e271. doi: 10.1161/CIRCULATIONAHA.110.017665
2. Rehman R, Yelamanchili VS, Makaryus AN. *Cardiac Imaging. 2017*. In: StatPearls Publishing; 2022.
3. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12:e1001779. doi: 10.1371/journal.pmed.1001779
4. Seetharam K, Shrestha S, Sengupta P. Artificial intelligence in cardiac imaging. *US Cardiology Review*. 2020;13:110–116.
5. Mitchell TM. *Machine learning. Vol. 1*. McGraw-hill New York; 1997.
6. Raisi-Estabragh Z, Salih A, Gkontra P, Atehortúa A, Radeva P, Boscolo Galazzo I, Menegaz G, Harvey NC, Lekadir K, Petersen SE. Estimation of biological heart age using cardiovascular magnetic resonance radiomics. *Sci Rep*. 2022;12:1–12. doi: 10.1038/s41598-022-16639-9
7. Owusu E, Boakye-Sekyerehene P, Appati JK, Ludu JY. Computer-aided diagnostics of heart disease risk prediction using boosting support vector machine. *Comput Intell Neurosci*. 2021;2021:1–12. doi: 10.1155/2021/3152618
8. Ali L, Niamat A, Khan JA, Golilarz NA, Xingzhong X, Noor A, Nour R, Bukhari SAC. An optimized stacked support vector machines based expert system for the effective prediction of heart failure. *IEEE Access*. 2019:54007–54014.
9. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–444. doi: 10.1038/nature14539
10. Lekadira K, Osuala R, Gallin CS, Lazrak N, Kushibar K, Tsakou G, Auss'o S, Alberich LC, Marias K, Tskinakis M, et al. FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. *ArXiv*. Preprint posted online September 20, 2021; updated September 29, 2021. https://doi.org/10.48550/arXiv.2109.09658
11. Chen H, Gomez C, Huang CM, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. *NPG Digit Med*. 2022;5:1–15. doi: 10.1038/s41746-022-00699-2
12. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI-explainable artificial intelligence. *Sci Robot*. 2019;4:eaay7120. doi: 10.1126/scirobotics.aay7120
13. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy*. 2020;23:18. doi: 10.3390/e23010018
14. Galazzo IB, Cruciani F, Brusini L, Salih A, Radeva P, Storti SF, Menegaz G. Explainable artificial intelligence for magnetic resonance imaging aging brainprints: grounds and challenges. *IEEE Signal Process Mag*. 2022;39:99–116.
15. Bauer K, Hinz O, Aalst W van der, Weinhardt C. Expl (AI) n it to me–explainable AI and information systems research. *Bus Inf Syst Eng*. 2021;63:79–82.
16. 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, Italy; 2017:618–626. doi: 10.1109/ICCV.2017.74
17. Molnar C. Interpretable machine learning. Lulu. com, 2020.
18. Zhang Q, Hann E, Werys K, Wu C, Popescu I, Lukaschuk E, Barutcu A, Ferreira VM, Piechnik SK. Deep learning with attention supervision for automated motion artefact detection in quality control of cardiac T1-mapping. *Artif Intell Med*. 2020;110:101955. doi: 10.1016/j.artmed.2020.101955
19. Smilkov D, Thorat N, Kim B, Viégas FB, Wattenberg M. SmoothGrad: removing noise by adding noise. *ArXiv*. Preprint posted online June 12, 2017. https://doi.org/10.48550/arXiv.1706.03825
20. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016:2921–2929.
21. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Workshop at International Conference on Learning Representations. Citeseer; 2014.
22. Raisi-Estabragh Z, McCracken C, Gkontra P, Jaggi A, Ardissino M, Cooper J, Biasiolli L, Aung N, Piechnik SK, Neubauer S, et al. Associations of meat and fish consumption with conventional and radiomics cardiovascular magnetic resonance phenotypes in the UK Biobank. *Front Cardiovasc Med*. 2021;8:667849. doi: 10.3389/fcvm.2021.667849
23. Raisi-Estabragh Z, Izquierdo C, Campello VM, Martin-Isla C, Jaggi A, Harvey NC, Lekadir K, Petersen SE. Cardiac magnetic resonance radiomics: basic principles and clinical perspectives. *Eur Heart J Cardiovasc Imaging*. 2020;21:349–356. doi: 10.1093/ehjci/jeaa028
24. Van Griethuysen JJ, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, Beets-Tan RG, Fillion-Robin JC, Pieper S, Aerts HJ. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77:e104–e107. doi: 10.1158/0008-5472.can-17-0339
25. Greenwell BM, Boehmke BC, McCarthy AJ. A simple and effective model-based variable importance measure. *ArXiv*. Preprint posted online May 12, 2018. https://doi.org/10.48550/arXiv.1805.0475
26. Apley DW, Zhu J. Visualizing the effects of predictor variables in black box supervised learning models. *J R Stat Soc*. 2020;82:1059–1086.
27. Chauhan D, Anyanwu E, Goes J, Besser SA, Anand S, Madduri R, Getty N, Kelle S, Kawaji K, Mor-Avi V, et al. Comparison of machine learning and deep learning for view identification from cardiac magnetic resonance images. *Clin Imaging*. 2022;82:121–126. doi: 10.1016/j.clinimag.2021.11.013
28. Pérez-Pelegrió M, Monmeneu JV, López-Lereu MP, Pérez-Pelegrió L, Maceira AM, Bodió V, Moratal D. Automatic left ventricle

volume calculation with explainability through a deep learning weak-supervision methodology. *Comput Methods Programs Biomed*. 2021;208:106275. doi: 10.1016/j.cmpb.2021.106275
29. Biswas SK, Chakraborty M, Purkayastha B, Roy P, Thounaojam DM. Rule extraction from training data using neural network. *Int J Artif Intell Tools*. 2017;26:1750006.
30. Bondarenko A, Aleksejeva L, Jumutc V, Borisov A. Classification tree extraction from trained artificial neural networks. *Procedia Comput Sci*. 2017;104:556–563. doi: 10.1016/j.procs.2017.01.172
31. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:1448–1458.
32. Kucukseymen S, Arafati A, Al-Otaibi T, El-Rewaidy H, Fahmy AS, Ngo LH, Nezafat R. Noncontrast cardiac magnetic resonance imaging predictors of heart failure hospitalization in heart failure with preserved ejection fraction. *J Magn Reson Imaging*. 2022;55:1812–1825. doi: 10.1002/jmri.27932
33. Cau R, Pisu F, Porcu M, Cademartiri F, Montisci R, Bassareo P, Muscogiuri G, Amadu A, Sironi S, Esposito A, et al. Machine learning approach in diagnosing Takotsubo cardiomyopathy: the role of the combined evaluation of atrial and ventricular strain, and parametric mapping. *Int J Cardiol*. 2023;373:124–133. doi: 10.1016/j.ijcard.2022.11.021
34. Singh A, Kwiecinski J, Miller RJ, Otaki Y, Kavanagh PB, Van Kriekinge SD, Parekh T, Gransar H, Pieszko K, Killekar A, et al. Deep learning for explainable estimation of mortality risk from myocardial positron emission tomography images. *Circ Cardiovasc Imaging*. 2022;15:e014526. doi: 10.1161/CIRCIMAGING.122.014526
35. Lagopoulos A, Hristu-Varsakelis D. Measuring the left ventricular ejection fraction using geometric features. In: 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS). IEEE. 2022:1–6.
36. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016:1135–1144.
37. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*. 2015;10:e0130140. doi: 10.1371/journal.pone.0130140
38. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller MA. Striving for simplicity: the all convolutional net. *arXiv*. Preprint posted online December 21, 2014; updated April 13, 2015. https://doi.org/10.48550/arXiv.1412.6806
39. Tong Q, Li C, Si W, Liao X, Tong Y, Yuan Z, Heng PA. RIANet: recurrent interleaved attention network for cardiac MRI segmentation. *Comput Biol Med*. 2019;109:290–302. doi: 10.1016/j.compbiomed.2019.04.042
40. Nurmaini S, Partan RU, Bernolian N, Sapitri AI, Tutuko B, Rachmatullah MN, Darmawahyuni A, Firdaus F, Mose JC. Deep learning for improving the effectiveness of routine prenatal screening for major congenital heart diseases. *J Clin Med*. 2022;11:6454. doi: 10.3390/jcm11216454
41. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: International Conference on Machine Learning. PMLR. 2017:3145–3153.
42. Strobelt H, Gehrmann S, Behrisch M, Perer A, Pfister H, Rush AM. S eq 2s eq-v is: a visual debugging tool for sequence-to-sequence models. *IEEE Trans Vis Comput Graph*. 2018;25:353–363. doi: 10.1109/TVCG.2018.2865044
43. Ghorbani A, Ouyang D, Abid A, He B, Chen JH, Harrington RA, Liang DH, Ashley EA, Zou JY. Deep learning interpretation of echocardiograms. *NPJ Digit Med*. 2020;3:1–10. doi: 10.1038/s41746-019-0216-8
44. Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognit*. 2017;65:211–222. doi: 10.1016/j.patcog.2016.11.008
45. Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision. 2015:1520–1528.
46. Kindermans PJ, Schütt KT, Alber M, Müller KR, Erhan D, Kim B, Dähne S. Learning how to explain neural networks: PatternNet and PatternAttribution. *arXiv*. Preprint posted online May 16, 2017; updated October 24, 2017. https://doi.org/10.48550/arXiv.1705.05598
47. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: International Conference on Machine Learning. PMLR. 2017:3319–3328.
48. Vafaeezadeh M, Behnam H, Hosseinsabet A, Gifani P. Automatic morphological classification of mitral valve diseases in echocardiographic images based on explainable deep learning methods. *Int J Comput Assist Radiol Surg*. 2022;17:413–425. doi: 10.1007/s11548-021-02542-7
49. Ragnarsdottir H, Manduchi L, Michel H, Laumer F, Wellmann S, Ozkan E, Vogt JE. Interpretable prediction of pulmonary hypertension in newborns using echocardiograms. In: DAGM German Conference on Pattern Recognition. Springer; 2022:529–542.
50. Jiao Y, Yuan J, Sodimu OM, Qiang Y, Ding Y. Deep neural network-aided histopathological analysis of myocardial injury. *Front Cardiovasc Med*. 2021;8:724183. doi: 10.3389/fcvm.2021.724183
51. Wang Y, Chen W, Tang T, Xie W, Jiang Y, Zhang H, Zhou X, Yuan K. Cardiac segmentation method based on domain knowledge. *Ultrason Imaging*. 2022;44:105–117. doi: 10.1177/01617346221099435
52. Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian V. Grad-CAM++: improved visual explanations for deep convolutional networks. *ArXiv*. Preprint posted online October 30, 2017; updated November 9, 2018. https://doi.org/10.48550/arXiv.1710.11063
53. Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, et al. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: International Conference on Machine Learning. PMLR. 2018:2668–2677.
54. Clough JR, Oksuz I, Puyol-Antón E, Ruijsink B, King AP, Schnabel JA. Global and local interpretability for cardiac MRI classification. In: International Conference on Medical Image Computing and Computer Assisted Intervention. Springer; 2019:656–564.
55. Janik A, Dodd J, Ifrim G, Sankaran K, Curran K. Interpretability of a deep learning model in the application of cardiac MRI segmentation with an ACDC challenge dataset. In: Medical Imaging 2021: image processing. Vol. 11596. SPIE. 2021:861–872.
56. Guo F, Ng M, Goubran M, Petersen SE, Piechnik SK, Neubauer S, Wright G. Improving cardiac MRI convolutional neural network segmentation on small training datasets and dataset shift: a continuous kernel cut approach. *Med Image Anal*. 2020;61:101636. doi: 10.1016/j.media.2020.101636
57. Abdi AH, Luong C, Tsang T, Allan G, Nouranian S, Jue J, Hawley D, Fleming S, Gin K, Swift J, et al. Automatic quality assessment of echocardiograms using convolutional neural networks: feasibility on the apical four-chamber view. *IEEE Trans Med Imaging*. 2017;36:1221–1230. doi: 10.1109/TMI.2017.2690836
58. Otaki Y, Singh A, Kavanagh P, Miller RJ, Parekh T, Tamarappoo BK, Sharir T, Einstein AJ, Fish MB, Ruddy TD, et al. Clinical deployment of explainable artificial intelligence of SPECT for diagnosis of coronary artery disease. *JACC Cardiovasc Imaging*. 2022;15:1091–1102. doi: 10.1016/j.jcmg.2021.04.030
59. Chartsias A, Joyce T, Papanastasiou G, Semple S, Williams M, Newby DE, Dharmakumar R, Tsaftaris SA. Disentangled representation learning in cardiac image analysis. *Med Image Anal*. 2019;58:101535. doi: 10.1016/j.media.2019.101535
60. Seitzer M, Yang G, Schlemper J, Oktay O, Würfl T, Christlein V, Wong T, Mohiaddin R, Firmin D, Keegan J, et al. Adversarial and perceptual refinement for compressed sensing MRI reconstruction. In: International conference on medical image computing and computer-assisted intervention. Springer; 2018:232–240.
61. Danese MD, Gleeson M, Kutikova L, Griffiths RI, Azough A, Khunti K, Seshasai SRK, Ray KK. Estimating the economic burden of cardiovascular events in patients receiving lipid-modifying therapy in the UK. *BMJ Open*. 2016;6:e011805. doi: 10.1136/bmjopen-2016-011805
62. Walker S, Asaria M, Manca A, Palmer S, Gale CP, Shah AD, Abrams KR, Crowther M, Timmis A, Hemingway H, et al. Long-term healthcare use and costs in patients with stable coronary artery disease: a populationbased cohort using linked health records (CALIBER). *Eur Heart J Qual Care Clin Outcomes*. 2016;2:125–140. doi: 10.1093/ehjqcco/qcw003
63. Hamon R, Junklewitz H, Sanchez I, Malgieri G, De Hert P. Bridging the gap between AI and explainability in the GDPR: towards trustworthiness-by-design in automated decision-making. *IEEE Comput Intell Mag*. 2022;17:72–85. doi: 10.1109/mci.2021.3129960
64. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3:e745–e750. doi: 10.1016/s2589-7500(21)00208-9
65. Doshi-Velez F Kim B. Towards a rigorous science of interpretable machine learning. *ArXiv*. Preprint posted online February 28, 2017; updated March 2, 2017. https://doi.org/10.48550/arXiv.1702.08608
66. Samek W, Binder A, Montavon G, Lapuschkin S, Müller KR. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans Neural Networks Learn Syst*. 2017;28:2673. doi: 10.1109/TNNLS.2016.2599820
67. Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf Fusion*. 2021;76:89–106. doi: 10.1016/j.inffus.2021.05.009
68. Miller RJ, Kuronuma K, Singh A, Otaki Y, Hayes S, Chareonthaitawee P, Kavanagh P, Parekh T, Tamarappoo BK, Sharir T, et al. Explainable deep learning improves physician interpretation of myocardial perfusion imaging. *J Nucl Med*. 2022;63:1768–1774. doi: 10.2967/jnumed.121.263686

69. Tamarappoo BK, Lin A, Commandeur F, McElhinney PA, Cadet S, Goeller M, Razipour A, Chen X, Gransar H, Cantu S, et al. Machine learning integration of circulating and imaging biomarkers for explainable patientspecific prediction of cardiac events: a prospective study. *Atherosclerosis*. 2021;318:76–82. doi: 10.1016/j.atherosclerosis.2020.11.008
70. Wang J, Liu X, Wang F, Zheng L, Gao F, Zhang H, Zhang X, Xie W, Wang B. Automated interpretation of congenital heart disease from multi-view echocardiograms. *Med Image Anal*. 2021;69:101942. doi: 10.1016/j.media.2020.101942
71. Painchaud N, Duchateau N, Bernard O, Jodoin PM. Echocardiography segmentation with enforced temporal consistency. *IEEE Trans Med Imaging*. 2022;41:2867–2878. doi: 10.1109/TMI.2022.3173669
72. Puyol-Antón E, Chen C, Clough JR, Ruijsink B, Sidhu BS, Gould J, Porter B, Elliott M, Mehta V, Rueckert D, et al. Interpretable deep models for cardiac resynchronisation therapy response prediction. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. Springer; 2020:284–293.
73. Biffi C, Oktay O, Tarroni G, Bai W, De Marvao A, Doumou G, Rajchl M, Bedair R, Prasad S, Cook S, et al. Learning interpretable anatomical features through deep generative models: application to cardiac remodeling. In: medical image computing and computer assisted intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11. Springer; 2018:464–471.
74. Sakai A, Komatsu M, Komatsu R, Matsuoka R, Yasutomi S, Dozen A, Shozu K, Arakaki T, Machino H, Asada K, et al. Medical professional enhancement using explainable artificial intelligence in fetal cardiac ultrasound screening. *Biomedicines*. 2022;10:551. doi: 10.3390/biomedicines10030551
75. Sun J, Darbehani F, Zaidi M, Wang B. Saunet: shape attentive u-net for interpretable medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23. Springer; 2020:797–806.
76. Alabed S, Uthoff J, Zhou S, Garg P, Dwivedi K, Alandejani F, Gosling R, Schobs L, Brook M, Shahin Y, et al. Machine learning cardiac-MRI features predict mortality in newly diagnosed pulmonary arterial hypertension. *Eur Heart J Digit Health*. 2022;3:265–275. doi: 10.1093/ehjdh/ztac022
77. Singh A, Miller RJ, Otaki Y, Kavanagh P, Hauser MT, Tzolos E, Kwiecinski J, Van Kriekinge S, Wei CC, Sharir T, et al. Direct risk assessment from myocardial perfusion imaging using explainable deep learning. *JACC Cardiovasc Imaging*. 2023;16:209–220. doi: 10.1016/j.jcmg.2022.07.017
78. Valvano G, Leo A, Tsaftaris SA. Regularizing disentangled representations with anatomical temporal consistency. In: *Biomedical Image Synthesis and Simulation*. Elsevier; 2022:325–346.