Marey *et al. Egypt J Radiol Nucl Med (2024) 55:183*
https://doi.org/10.1186/s43055-024-01356-2

Egyptian Journal of Radiology
and Nuclear Medicine

REVIEW **Open Access**

Explainability, transparency and black box challenges of AI in radiology: impact on patient care in cardiovascular radiology

Ahmed Marey[1*], Parisa Arjmand[2], Ameerh Dana Sabe Alerab[3], Mohammad Javad Eslami[4], Abdelrahman M. Saad[1], Nicole Sanchez[5,6] and Muhammad Umair[7]

Abstract

The integration of artificial intelligence (AI) in cardiovascular imaging has revolutionized the field, offering significant advancements in diagnostic accuracy and clinical efficiency. However, the complexity and opacity of AI models, particularly those involving machine learning (ML) and deep learning (DL), raise critical legal and ethical concerns due to their "black box" nature. This manuscript addresses these concerns by providing a comprehensive review of AI technologies in cardiovascular imaging, focusing on the challenges and implications of the black box phenomenon. We begin by outlining the foundational concepts of AI, including ML and DL, and their applications in cardiovascular imaging. The manuscript delves into the "black box" issue, highlighting the difficulty in understanding and explaining AI decision-making processes. This lack of transparency poses significant challenges for clinical acceptance and ethical deployment. The discussion then extends to the legal and ethical implications of AI's opacity. The need for explicable AI systems is underscored, with an emphasis on the ethical principles of beneficence and non-maleficence. The manuscript explores potential solutions such as explainable AI (XAI) techniques, which aim to provide insights into AI decision-making without sacrificing performance. Moreover, the impact of AI explainability on clinical decision-making and patient outcomes is examined. The manuscript argues for the development of hybrid models that combine interpretability with the advanced capabilities of black box systems. It also advocates for enhanced education and training programs for healthcare professionals to equip them with the necessary skills to utilize AI effectively. Patient involvement and informed consent are identified as critical components for the ethical deployment of AI in healthcare. Strategies for improving patient understanding and engagement with AI technologies are discussed, emphasizing the importance of transparent communication and education. Finally, the manuscript calls for the establishment of standardized regulatory frameworks and policies to address the unique challenges posed by AI in healthcare. By fostering interdisciplinary collaboration and continuous monitoring, the medical community can ensure the responsible integration of AI into cardiovascular imaging, ultimately enhancing patient care and clinical outcomes.

Keywords Artificial intelligence, Cardiovascular imaging, Machine learning, Deep learning, Black box phenomenon, Explainable AI, Ethical implications, Clinical decision-making, Patient outcomes, Regulatory frameworks

*Correspondence:
Ahmed Marey
ahmed.ahmed1797@alexmed.edu.eg
Full list of author information is available at the end of the article

Introduction

The pervasive integration of AI in cardiovascular imaging raises pertinent legal and ethical considerations. The increasing complexity of AI models, particularly those involving machine learning (ML) and deep learning (DL), introduces the challenge of the "black box"—a term that describes the opacity of AI decision-making processes [1]. This challenge raised concerns about the explicability and transparency of AI systems, which are essential for their ethical deployment and clinical acceptance [2].

Despite numerous advancements in AI-driven cardiovascular imaging, existing reviews often focus predominantly on technical enhancements and clinical outcomes, with less emphasis on how these AI models make decisions or on the mechanisms underlying their outputs [3]. This gap underscores a vital need for comprehensive reviews that not only explores these advanced technologies but also delves into the ethical and practical implications of their opaque nature.

This manuscript is structured to first outline the basic concepts and technologies underpinning AI in cardiovascular imaging, followed by an exploration of the "black box" phenomenon. Subsequent sections discuss legal, ethical, and practical challenges, culminating in a discussion on future directions that bridge gaps between technical capabilities and clinical needs. Our goal is to furnish clinicians, researchers, and policymakers with a deeper understanding of AI's potential and limitations in cardiovascular healthcare.

An overview on AI

AI involves developing computer programs that perform complex tasks mimicking human cognition. A key component of AI, machine learning (ML), enables algorithms to learn from data, improve performance, and make predictions [4]. Advances in computational power and big data have propelled ML's application in healthcare [5]. The rise of smart devices and electronic medical records has expanded data availability, enhancing ML algorithm performance despite data complexity [6].

ML training may be either "supervised" or "unsupervised." In supervised training, an ML model is trained on a range of inputs in association with a known outcome which is supervised, either in accordance to an objective classification metric or by a domain expert. In contrast, unsupervised training refers to the development of a model to explore the patterns or clusters that are not well-defined inside datasets. In this form, the model is only provided by unlabeled input data and does not learn to fit data to an outcome [7].

Deep learning (DL), a subset of ML, is another crucial concept in AI. DL is programmed to process data with large artificial neural networks through multiple processing layers, resembling the working of biological neurons [8]. It has achieved impressive results when used for complex tasks involving very high-dimensional data, including speech and image recognition to self-driving cars [9, 10]. Deep learning models utilize numerous layers of hidden neurons to generate increasingly abstract and nonlinear representations of the underlying data. This process, known as "representation learning," constitutes a pivotal aspect of deep neural networks. Following the acquisition of these representations, final output nodes are frequently utilized as inputs for logistic regression models or support vector machines (SVMs) to perform the ultimate regression or classification tasks. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) represent two prominent forms of deep learning models for supervised learning. The primary distinction between CNNs and RNNs lies in their respective layer designs. Beyond these methods, there exists a diverse array of deep neural network architectures.

CNNs resemble fully connected neural networks, comprising neurons with adjustable weights and biases. Their potency stems from the capacity to establish local connectivity across images or signals. These localized connections incorporate nonlinear activation functions, facilitating the transformation of representations into higher, slightly more abstract forms. Furthermore, shared weights across layers, layer pooling, and the integration of numerous hidden layers enable the learning of highly intricate functions. In contrast, RNNs excel in processing sequential data such as speech and language. Comprised of an additional hidden state vector, RNNs retain "memory" regarding the historical data observations, rendering them well-suited for tasks involving sequential information [11, 12].

In recent years, generative AI (GAI), a subtype of AI, peaked with the introduction of new language and image models that showed unprecedented capabilities. GAI models can now create images or even videos from text input, edit images from text prompts, and generate text taking part in complete conversations. These models have also openly available feeding more fuel into the surge of GAI's popularity among the general public. Hence, this allowed non-technical users to experiment with use cases in various domains and specialties. GAI can be routed back to the advancement of specifically two type of networks: transformers, which are more complex forms of RNNs and GANs which use two different CNNs and train them together in an adversarial manner.

While RNNs excel at handling sequential data by maintaining a hidden state that captures the essence of previous inputs, they struggle with long-range dependencies and parallel processing. The transformer overcomes these limitations by introducing self-attention mechanisms, which allow the model to weigh the importance of each word in a sequence relative to all other words, rather than relying solely on the sequential processing inherent in RNNs. The transformer's architecture eliminates the need for recurrent connections, enabling it to process all tokens in a sequence simultaneously. This parallelism significantly enhances efficiency and allows the model to capture long-term dependencies more effectively. The use of multi-head self-attention within the transformer ensures that the model can focus on different parts of the sequence simultaneously, leading to a richer and more nuanced understanding of context. Transformers paved the way to the remarkable power and admirable status of Large Language Models (LLMs) like ChatGPT today [13].

Generative Adversarial Networks (GANs), on the other hand, are highly effective in generating realistic data across various domains, such as images, video, and audio. GANs consist of two neural networks—a generator and a discriminator—that are trained together in a competitive setting. The generator attempts to produce data that mimics the real data distribution; while, the discriminator tries to distinguish between real and generated data. This adversarial process pushes the generator to create increasingly convincing outputs, ultimately resulting in the generation of highly realistic data [14].

Teaching Points:

- AI and ML are critical for performing tasks that mimic human cognition, with ML enabling algorithms to learn from data and improve predictions.
- Deep Learning, a subset of ML, uses large neural networks to process complex data, excelling in tasks like image and speech recognition.
- GAI models, powered by transformers and GANs, are revolutionizing AI applications by creating realistic data, such as images and text, from simple inputs.

AI applications in cardiovascular imaging

AI can analyze vast amounts of image data to identify subtle patterns and anomalies that may be overlooked by human experts. For instance, AI-powered systems can accurately quantify coronary artery stenosis from CT angiography in real time [15]. Neural networks can also be trained with the appropriate data to detect early signs of heart failure from chest X-rays [16]. Such applications can lead to earlier and more accurate diagnoses, enabling timely interventions and improved patient outcomes. Beyond diagnostic capabilities, AI is optimizing imaging workflows. Automated image acquisition, reconstruction, and segmentation tasks reduce human error and expedite the interpretation process [17]. Additionally, AI-driven predictive models can identify patients at high risk for cardiovascular events based on imaging data, allowing for proactive risk management strategies [18].

Generative AI (GAI) is revolutionizing cardiovascular imaging by enhancing image quality, automating complex tasks, and improving diagnostic precision across various modalities [19].In Cardiac MRI (CMR), for example, GAI plays a crucial role in accelerating image reconstruction and reducing motion artifacts, with methods like those developed by Ghodrati et al. enabling free-breathing scans, thus enhancing patient comfort and scan efficiency [20]. Advanced reconstruction techniques such as variational neural networks (VNNs) allow for high-quality imaging from undersampled data, significantly reducing scan times without compromising accuracy [21]. This is particularly beneficial for procedures requiring detailed volumetric and functional analysis of the heart, making CMR more accessible and reliable for clinical decision-making.

In Cardiac Computed Tomography (CCT), GAI-based approaches have shown significant promise in improving both image quality and diagnostic accuracy. AI-driven algorithms, such as Itu et al.'s method for Fractional Flow Reserve CT (FFR-CT), have drastically reduced analysis time while maintaining high predictive accuracy, showcasing the potential of AI to enhance non-invasive coronary artery disease (CAD) evaluation [22]. These AI-powered advancements are not only streamlining clinical workflows but also providing more consistent and reliable diagnostic information, ultimately improving patient outcomes in cardiovascular care. Table 1 provides overview of the AI concepts and applications discussed in this section.

Teaching Points:

- AI improves diagnostic accuracy by identifying subtle patterns in cardiovascular imaging, such as detecting coronary artery stenosis and early heart failure.
- AI optimizes imaging workflows by automating tasks like image acquisition and reconstruction, reducing human error and speeding up diagnosis.

The "black box" challenges

The "black box" nature in AI models

Despite these benefits, the complexity of AI models, particularly deep learning methods, poses significant challenges [23, 24]. In the context of AI in radiology, "Black box" refers to situations where the AI model's decision-making process is opaque or not easily understandable by humans. This means that while the AI can provide results or recommendations, the underlying reasoning or mechanisms that led to these conclusions are not transparent. Such indications can pose challenges in clinical settings because clinicians may not fully understand or trust the

Table 1 AI Concepts and Applications

Concept	Key points	Applications
AI and ML in healthcare	AI mimics human cognition	Broad application in healthcare for complex tasks
	ML enables algorithms to learn and predict from data	Used for pattern recognition, prediction, and diagnosis
ML training types	Supervised: Trained with known outcomes	Diagnostic tool training, outcome prediction
	Unsupervised: Finds patterns in unlabeled data	Clustering patient data, anomaly detection
Deep learning (DL)	Utilizes large neural networks	Image and speech recognition, autonomous systems
	Representation learning through multiple layers	Enhanced diagnostic accuracy, particularly in imaging
DL models	CNNs: Local connectivity in images, nonlinear activation	Image analysis, feature extraction in medical imaging
	RNNs: Sequential data processing with memory	Natural language processing, time-series data analysis
Generative AI (GAI)	Creation of images, text, videos	Enhancing diagnostic imaging, automated content generation
Transformers	Self-attention mechanisms, parallel processing	Large Language Models (e.g., ChatGPT), text analysis, complex tasks
GANs	Adversarial training for realistic data generation	Image and video generation, audio synthesis, anomaly detection
AI in cardiovascular imaging	Early detection of diseases, workflow optimization, and risk prediction	CT angiography for coronary artery stenosis, heart failure detection
	GAI in CMR and CCT for improved image quality and diagnostic accuracy	Accelerated scan times, non-invasive coronary artery disease evaluation

AI's outputs, which can impact patient care [25, 26]. Understanding and explaining AI's decisions is crucial for clinical acceptance and ethical deployment [27].

Ensuring the reliability of an AI system requires demonstrating that the system has learned the underlying properties and that the decisions made are not based on irrelevant correlations between input and output values in the training dataset [28]. While it is possible to minimize an AI method's weaknesses by carefully selecting its model architecture and training algorithm, errors cannot be eliminated [29].

The ability of different AI models to understand generated models varies significantly. With the emergence of new and powerful DL methods, it is becoming increasingly difficult to reconstruct decisions. Frequently, the resulting models function as "black boxes," rendering it arduous for users to comprehend the internal processes [28]. Users can only understand input and output values, despite designers possessing an understanding of the system's architecture and the methodologies employed to generate the models [30]. In contrast, interpretable models are referred to as white boxes, where weights are assigned to each feature, allowing for easy reading and interpretation while an intermediate stage between the two is the gray box. Gray box models provide a certain level of insight into internal data processing [31].

It is important to note that in practice, a method cannot always be clearly classified as a white, gray, or black box method. Thus, to address the issue of lack of explainability, there is a need for explanation models for black box models, which help in understanding how they work. Figure 1 provides visual representation of the black box problem in comparison with explainable AI.

Challenges and limitations associated with AI's "Black-Box" nature in cardiovascular imaging

As stated before, the decision-making process of AI is often unclear, which presents a challenge in interpreting and understanding its results. Although the results of DL in cardiovascular imaging are promising, they

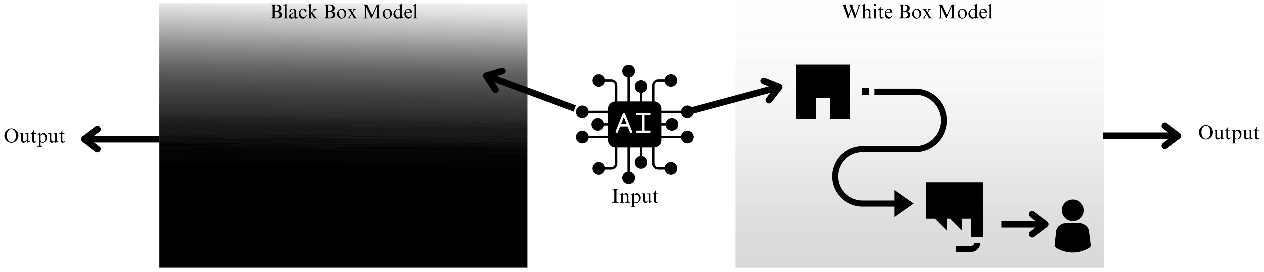

Fig. 1 Visual representation of the black box problem in comparison with explainable AI

are still modest, and several challenges must be overcome to improve them [32]. Common deep learning architectures, such as convolutional neural networks (CNNs), generative adversarial networks (GANs), and recurrent neural networks (RNNs) do not provide explanations for their outcomes [33]. In the clinical context, the most important challenge is often referred to as a black box. Therefore, in the health sector, developing explainable machine learning systems remains a top priority for computer scientists, policymakers, and users [34].

There is currently no agreed-upon definition for explainability, despite the consensus on the importance of developing and implementing interpretable models [35]. For instance, Luo et al. proposed a new data preprocessing technique for detecting cardiac diseases using cardiac magnetic resonance (CMR) imaging and a new network structure for the estimation of left ventricular volume. Their study demonstrated that the method had high accuracy in predicting left ventricular (LV) volumes. However, they pointed out a significant challenge commonly encountered in deep learning methods—the lack of interpretability for physicians. Achieving true interpretability in LV volume prediction may for example mean enabling physicians to identify the specific pixels used in blood volume computations. They emphasized that future research should focus on achieving interpretability in the direct prediction of LV volumes [36].

Even though AI algorithms can detect coronary artery disease, heart failure, conduction abnormalities, and valvular heart disease and aid in diagnoses, the lack of transparency raises concerns about their reliability, interpretability, and potential biases. To ensure that AI's clinical integration aligns with practical standards in healthcare, it is essential to understand the inner workings of these algorithms.

Teaching Points:

- The complexity of AI models, particularly deep learning methods, often leads to a "black box" phenomenon, where the decision-making process is not transparent and difficult to interpret.
- Explainability is crucial for clinical acceptance and ethical deployment of AI in healthcare, yet remains a challenge due to the opaque nature of many AI models.
- In cardiovascular imaging, the lack of transparency in AI models raises concerns about reliability, interpretability, and potential biases, making explainable AI a priority.

Impact of explainability on clinical decision-making and patients' outcomes in cardiovascular imaging AI

As elaborated before, in cardiovascular imaging, AI has an essential role, and understanding how it works is essential for effective implementation [37]. Evidence-based medicine is challenged by the opaqueness of ML models, especially in medical imaging. In evidence-based medicine, clinical decisions are informed by the best available evidence from scientific research, combined with clinical expertise and patient values. This approach relies heavily on transparent and interpretable data and models, allowing clinicians to understand the rationale behind recommendations or decisions. However, ML models, including those used in CV imaging, often operate as "black boxes," meaning their internal decision-making processes are not easily interpretable or explainable. This lack of transparency poses a significant challenge for evidence-based medicine because clinicians may struggle to trust or understand the outputs of these models, hindering their ability to integrate them effectively into clinical practice. In the context of CVS imaging, where accurate diagnosis and treatment decisions are paramount, the opaqueness of ML models can lead to uncertainty or skepticism among healthcare professionals. Clinicians may hesitate to rely on ML-based recommendations without a clear understanding of how the model arrived at its conclusions.

One of the significant challenge is related to error detection. It is plausible that AI systems may sometimes deviate from accepted standards of clinical decision-making [38]. Image classification algorithms, such as convolutional neural networks, are particularly susceptible to unexpected and unusual classification errors [39], leading to difficulty in comprehending the causal factors influencing these ML models' correlations. This ambiguity can undermine healthcare practitioners' confidence in relying on AI predictions, particularly when they conflict with conventional clinical judgment [40]. To optimize ML systems, it's imperative to comprehend their decision-making process. AI explainability allows individuals to understand how an AI model makes decisions, going beyond just improving AI actions [41].

Qualitative research indicates that clinicians prioritize pertinent and easily comprehensible ML model information to make informed decisions. A study conducted by Tonekaboni et al. found that clinicians do not necessarily prefer to understand the causal mechanisms of action behind ML decision-making. Instead, they prefer easily understandable and relevant information about how the model works in the context of healthcare decision-making. This information may include confidence scores, the reasoning behind a decision, and details that are tailored to the specific clinical context [27].

Lang and colleagues also have pointed out that some of the most effective applications of AI in cardiovascular imaging may not be explainable. This has raised concerns among some experts who suggest that the use of unexplainable models should be stopped due to the significant problems they may pose [38, 42]. In conclusion, while technical experts may not possess comprehensive understanding of machine learning (ML) algorithms, it is imperative that these systems furnish outputs or associated information enabling users to assess predictions pertinent to their clinical decision-making. Although efforts are underway to develop mechanisms for contextualizing ML predictions based on user needs, achieving full comprehension of AI predictions remains an evolving research frontier [43]. Table 2 provides and overview of the importance of.

Teaching Points.

- The lack of transparency in AI models, particularly in cardiovascular imaging, poses a challenge to evidence-based medicine by making it difficult for clinicians to understand and trust AI-generated recommendations.
- The opaqueness of AI models can undermine healthcare professionals' confidence, especially when AI predictions conflict with traditional clinical judgment.
- Clinicians prioritize AI outputs that are relevant, easily comprehensible, and tailored to specific clinical contexts, even if they do not fully understand the underlying mechanisms.

Legal and ethical implications

Challenges related to unexplainable AI in healthcare

The opacity in AI systems introduces significant legal and ethical challenges in healthcare. Clinician trust is crucial for AI integration into clinical workflows. A lack of explainability and transparency can lead to ethical dilemmas and affect reliance on AI for patient care [44]. Ethical principles such as beneficence (acting in the best interest of patients) and non-maleficence (do no harm) come into play when considering the potential risks associated with using AI systems with opaque decision-making processes. Transparency in algorithmic processes is key to facilitating comprehension [45]. In clinical settings, AI techniques must provide justifications for their decisions to increase clinicians' confidence in the accuracy of the results [46]. The use of AI models with low transparency or interpretability also raises concerns about accountability, patient safety, and decision-making processes. From a legal perspective, the issue of clinician trust intersects with liability and accountability. If clinicians rely on AI-driven diagnoses or treatment recommendations without fully understanding the rationale behind them, it can complicate matters in cases of medical errors or adverse outcomes. Determining responsibility becomes challenging when the decision-making process of AI remains opaque, potentially raising questions about liability and legal accountability [38].

Unfortunately, many AI-based cardiovascular imaging applications often exhibit an unexplainable "black box." It can be challenging to evaluate the clinical risks and benefits of unexplainable models, particularly when there is a risk of biased decision-making. The challenge becomes even greater when it comes to distinguishing between AI models that can be explained and those that cannot [30]. The use of unexplainable AI in medical applications has been a topic of debate in recent times. While some argue that regulations should deal more strictly with the unexplainable models, others believe that stricter regulations might impede innovation, clinical adoption, and lead to suboptimal patient outcomes [38]. The replication of clinical trials for technically unexplained models is uniquely challenging since commercial developers often do not wish to divulge their trade secrets [47]. Nevertheless, it is essential to recognize that the uncertainty surrounding medical interventions is not a new challenge. However, it is essential to recognize that the unique complexities of AI-based cardiovascular imaging applications warrant careful consideration of whether distinct regulatory approaches are necessary. This includes adherence

Table 2 Explainability in AI for cardiovascular imaging

Concept	Key points	Challenges
Explainability in evidence-based medicine	Lack of transparency in AI models hinders evidence-based clinical decision-making	Clinicians struggle to trust and integrate AI recommendations
Clinical confidence	The opaqueness of AI can lead to uncertainty and skepticism among healthcare professionals	Difficulty in relying on AI predictions when they conflict with clinical judgment
Relevant and understandable outputs	Clinicians prefer AI outputs that are easily interpretable and relevant to clinical contexts	Need for AI systems to provide information that supports clinical decision-making without full model comprehension

to validation plans and regulations set forth by regulatory bodies such as the FDA for the deployment of medical AI. [30].

Legal frameworks governing unexplainable AI extend to medical malpractice, making it more difficult for clinicians to set standards of care. The changing landscape necessitates a reevaluation of professional expectations and guidelines. Increasingly, AI-based care poses challenges to traditional ethical beliefs, as automated decision-making impacts comprehensibility [38, 48].

Another challenge in the context of unexplainable AI is the concept of informed consent. Clinical experts believe that informed consent is essential before using AI on patients. They also believe that computer-aided detection applications should be disclosed in reports, explaining the reasons for eventual disagreement. The provision of inaccurate information to patients and clinicians about the risks of AI algorithms may indeed constitute a breach of the duty of care, so the adequacy of information provided to users is crucial in making judgments. When it comes to information, however, they wonder what exactly needs to be disclosed to the patient [49]. These challenges become more sophisticated in the use of unexplainable AI. Patients have the right to understand and agree to the procedures or treatments suggested by AI algorithms.

Some proposals are made to avoid some legal and ethical issues: one possible solution is to efficiently extract interpretable features for disease classification by leveraging the power of deep learning. Researchers proposed techniques for extracting features from deep learning models that are not only accurate for disease classification but also interpretable by healthcare professionals. By leveraging the capabilities of deep learning algorithms, these techniques aim to identify and extract meaningful and interpretable features or patterns from medical images that are indicative of specific diseases or conditions [50]. This approach allows clinicians to better understand how the deep learning model arrives at its predictions by providing insights into the features or characteristics of the medical images that contribute to the classification process.

Another approach is to provide visible explanations of the output of neural networks after their application to medical images. GRADCAM, short for Gradient-weighted Class Activation Mapping, is a technique used in computer vision and deep learning for visualizing and understanding the decision-making process of convolutional neural networks (CNNs). It works by generating a heatmap that highlights the regions of an input image that are most important for CNN's classification decision. This heatmap is produced by computing the gradient of the predicted class score with respect to the final convolutional layer of the CNN. By visualizing which parts of the input image contribute most strongly to the network's decision, GRADCAM provides valuable insights into how the model is processing the data and making predictions. This can significantly improve the understanding of the decisions made by these networks and enhance the trust and adoption of AI technologies among medical professionals [45]. An example of GRADCAM use in a cardiovascular context was highlighted by Zhang et al., where they employed attention supervision in a deep learning model to guide a multi-stream Convolutional Neural Network (CNN) to focus on specific myocardial segments for automated motion artifact detection in cardiac T1-mapping [51]. However, some commentators have suggested it may be necessary to abandon unexplainable AI models. This is due to the significant problems that arise from the use of such models, which may be difficult to explain or understand [47].

European and American Multi-society Statement highlights numerous AI-related ethical challenges and opportunities. Recognizing the need for practical guidelines, a framework has been called for to assist AI practitioners. However, it's worth noting that the rapid pace of change in AI techniques and tools makes it challenging to maintain a comprehensive and up-to-date understanding of the ethical landscape [52, 53].

Physician liability and fault

The use of unexplainable AI models in cardiovascular imaging raises complex questions regarding physician liability within the existing medical malpractice framework. The foundation of medical practice is based on the duty of care, which includes providing treatment, information, follow-up, and maintaining patient confidentiality. However, the evolving landscape of AI in clinical settings introduces uncertainties regarding the appropriate standard of care for clinicians employing unexplainable models [38, 54].

At present, regulations do not appear to conceive of any legally significant distinction between medical imaging AI models that can be explained and those that cannot, leaving the question open as to whether this regulatory approach appropriately balances patient interests, and whether it strikes a balance between innovation and safety.

As establishing a direct link between breach of duty and patient harm becomes increasingly difficult in AI-related medical malpractice, causation becomes especially intricate. In cases of unexplainable AI models contributing to patient injury, true causation, determined by a "but-for" test, may prove elusive [55]. A legal cause-and-effect analysis adds to the complexity, especially with models that operate beyond human comprehension and are technically unexplainable [38]. It is difficult to hold physicians legally responsible for injury under circumstances where foreseeable outcomes are difficult to identify.

Physicians must provide treatment consistent with professional best practices as mandated by law. When someone claims medical malpractice, they must prove that a physician failed to meet their duty of care and that as a result, they suffered legally recognizable harm. Courts face considerable challenges when it comes to adopting perspectives about the unexplainable nature of AI models in medical imaging, potentially complicating the attribution of liability [56]. Establishing standards of care and legal causation in medical malpractice cases is a complex task and presents inherent difficulties. Furthermore, the introduction of unexplainable medical AI adds another layer of complexity in product liability cases, leading to discussions on whether manufacturers should be held accountable for the unforeseeable outcomes of their products [56]. For instance, if a DL-powered model is used for cardiovascular CT image reconstruction and a patient is injured due to misdiagnosis of a cardiovascular abnormality, it may not be immediately clear whether the physician is responsible for the injury, even if a court finds that the physician breached their duty of care. Notably, the automatic presumption of fault in product liability regimes contrasts with the evidence-based approach in civil liability regimes [45].

Currently, the European and North American Multisociety Statement mentioned that physicians, including radiologists, are held liable in cases where "standard of care" is not provided. In cases where Al is used as a decision aid, radiologists will likely still be considered liable, though it is probable that litigation will also accuse Al product manufacturers. Since models incorporate large amounts of data, some of which are not perceptible to humans, the question will arise whether physicians should remain solely responsible or whether responsibility should be shifted to those who produce, market, and sell models. If, for example, low-dose CT images are enhanced by an algorithm to improve image quality, but this processing alters an important, but subtle feature so much that it is barely perceptible, the software developer should be liable for this. In the end, it is up to practice and case law to resolve these complex legal issues [52]. The American College of Radiology (ACR) also believes that, for now, since there are no diagnostic radiology models cleared for autonomous use in the U.S., radiologist responsibility remains solely with them.

Ultimately, the evolving landscape of medical AI necessitates careful consideration of regulatory approaches, ongoing technological advancements, and dynamic interpretations by courts. It is still difficult to strike a balance between encouraging innovation, ensuring patient safety, and setting clear standards of accountability. Despite existing literature discussing multiple liability theories about AI use, a definitive and unanimous answer to this issue has not yet been found [49]. In the coming years, solutions that improve interpretability and transparency while considering ethical considerations will play a pivotal role in shaping the responsible integration of AI into cardiovascular imaging. Table 3 provides a brief overview of the ethical and legal implications of unexplainable AI in CVS imaging.

Teaching Points:

- The opacity of AI systems raises significant ethical and legal challenges, particularly regarding accountability in cases of medical errors or adverse outcomes. Clinicians need transparency to build trust and make informed decisions.
- The use of unexplainable AI models complicates physician liability within the existing medical malpractice framework. Establishing a clear standard of care and causation becomes increasingly difficult when AI decisions are not fully understood.
- Informed consent is crucial when using AI in healthcare. Patients have the right to understand and agree to AI-driven procedures or treatments, making transparency in AI systems vital for maintaining trust and ensuring patient safety.

Bridging the gap and future directions

To overcome the ethical issues and challenges associated with the use of unexplainable AI in healthcare, particularly in cardiovascular imaging, there has been a surge of interest in explainable AI (XAI) techniques. This section

Table 3 Legal and ethical implications of unexplainable AI in cardiovascular imaging

Concept	Key points	Challenges
Transparency and accountability	Opacity in AI models leads to ethical and legal concerns regarding accountability in patient care	Difficulty in assigning liability when AI decisions are opaque
Physician liability	Unexplainable AI complicates liability issues within the medical malpractice framework	Challenges in establishing a standard of care and causation
Informed consent and patient safety	Transparency is essential for informed consent and maintaining patient trust in AI-driven care	Ensuring patients and clinicians understand the risks and benefits of AI

explores these techniques and outlines future directions to address the black box problem.

Advancements in explainable AI (XAI) Techniques and innovative solutions for interpretability

Model-based versus post hoc explanation

Model-based explanation refers to models, including linear regression or support vector machines that are simple enough to be easily understood while still being sophisticated enough to effectively capture the relationship between inputs and output [43]. These models are usually the traditional ML models that are simpler and more interpretable, in contrast to more modern complex models such as deep neural networks. Sparsity and simulatability are two well-known examples of these models. Sparsity refers to models that force many coefficients exactly to zero. Hence, this leads to a sparse model where only a subset of features significantly contributes to output, making the inner construct of this model explainable [57] Simulatability implies whether a human can internally reason about the model's computations and decision-making process. In simpler models, such as linear regression, it's easier for an individual to comprehend how each feature contributes to the final output [58]. Figure 2 shows how some explainable models can have minimal black box problem.

In contrast to model-based explanation, post hoc explanation trains a neural network and subsequently tries to elucidate the behavior of the resulting black box network rather than forcing the neural network to be explainable. This makes the post hoc to be easier to understand and more user-friendly and can be applied to any model, regardless of its complexity [57]. Techniques include inspection of the learned features, feature importance, interaction of features, and visual explanation by saliency maps [59–62]. However, the weakness of this method is its limited capacity to capture the full complexity of a model. Therefore, the

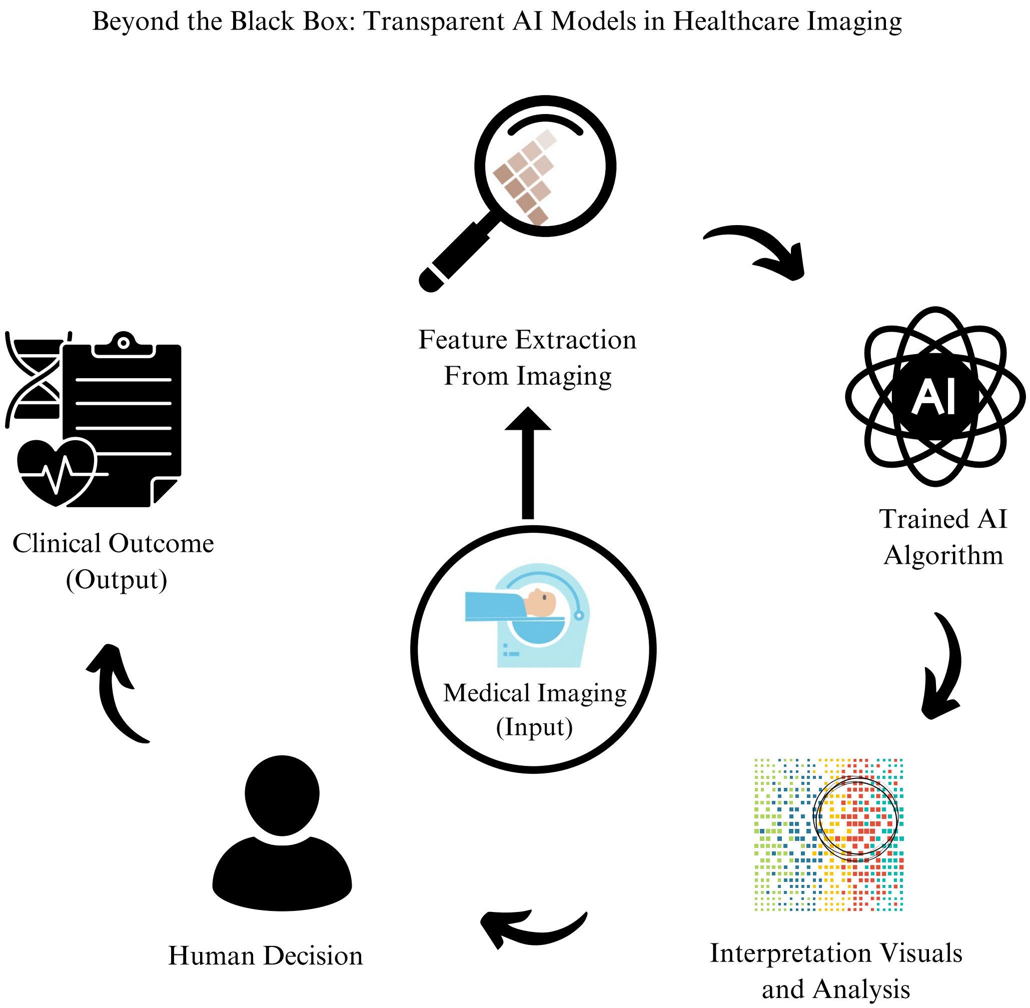
Beyond the Black Box: Transparent AI Models in Healthcare Imaging
Feature Extraction From Imaging
AI
Trained AI Algorithm
Interpretation Visuals and Analysis
Medical Imaging (Input)
Human Decision
Clinical Outcome (Output)

Fig. 2 An illustration of how some explainable models can have minimal black box problem

choice between these two is a trade-off between accuracy and interpretability and depends on the specific case used.

The global and local explanation

Global explanation, also called dataset-level explanation, refers to understanding the overall workings of a machine learning model across the entire dataset. It can quantify the importance of features and present them as scores at the dataset level. In this way, it is determined how much the features contribute to the output in the entire data set [60]. Local explanation explains how the model reached a particular decision, in every instance or data point. As an example, in a neural network model, the global explanation can find out at the "dataset level" that high blood pressure can increase the risk of cardiovascular events. While the local explanation shows why an increase in blood pressure leads to an increase in the risk of cardiovascular events in "a single person" [57, 63].

There are examples of global and local explainability in cardiovascular imaging as well. In 2019, Clough et al. presented a classification framework for identifying cardiac diseases using temporal sequences of cardiac MR segmentation based on a convolutional neural networks (CNN) model [64]. Their model not only performed the classification but, with the help of variational autoencoders (VAE), also allowed global and local interpretation. Variational autoencoders (VAEs) are a type of generative model that learns a latent representation of input data and can reconstruct input data from a compressed latent space [65]. By local interpretation, they meant the ability to ask, “Which features of this particular image led to it being classified in this particular way?” and by global interpretation, they meant, “Which common features were generally associated with images assigned to this particular class?”.

Techniques for interpretable features and visual explanations

Techniques that extract interpretable features from deep learning models are essential for demystifying black box AI systems. Research should focus on developing methods that transform complex neural network representations into more understandable formats without losing the accuracy and robustness of the original models [45]. Furthermore, visual explanation tools such as Gradient-weighted Class Activation Mapping (GRADCAM) can provide intuitive insights into AI decisions by highlighting important regions in an image that contribute to the model’s output. These visual aids can help clinicians understand and trust AI diagnoses by showing which parts of an image were most influential [62].

Development of hybrid models

Hybrid models that combine interpretable models with black box systems can enhance transparency without sacrificing performance. These models can use interpretable components to provide explanations and black box components to handle complex, high-dimensional data [66]. Research should focus on optimizing these hybrid approaches to maintain accuracy while improving interpretability.

User-friendly interfaces

Also, designing user-friendly interfaces that present AI explanations in an accessible manner is crucial. Future research should prioritize developing tools and platforms that allow clinicians to interact with and query AI models easily. Interactive dashboards, visualization tools, and customizable explanation reports can help make AI insights more usable and trustworthy [67].

By advancing XAI techniques and developing innovative solutions for interpretability, the medical community can enhance the transparency and trustworthiness of AI models in cardiovascular imaging. These efforts will facilitate the responsible and effective integration of AI technologies into clinical practice, ultimately leading to better patient outcomes and improved healthcare delivery.

Education and training for healthcare professionals

The integration of AI in healthcare, particularly in cardiovascular imaging, necessitates comprehensive education and training programs for healthcare professionals. These programs are essential for equipping clinicians with the necessary skills to understand, interpret, and effectively use AI models in their practice. Without proper training, the benefits of AI cannot be fully realized, and the potential for misuse or mistrust may increase [27].

Training programs should be designed to provide a robust understanding of AI concepts, including machine learning, deep learning, and explainable AI (XAI) [68]. These programs should cover both the theoretical foundations and practical applications of AI in healthcare. Clinicians need to understand not only how to use AI tools but also how these tools work, their limitations, and the ethical considerations involved [69].

An example of successful AI training can be found in radiology. Many radiology departments have started incorporating AI training into their residency programs. These programs often include courses on AI fundamentals, hands-on training with AI tools, and case studies demonstrating AI applications in radiological practice [70]. For instance, the Radiological Society of North America (RSNA) offers educational resources and workshops on AI, helping radiologists stay updated with the latest AI advancements and best practices [71].

Workshops and continuing education programs (CMEs) are vital for keeping healthcare professionals abreast of the latest developments in AI. Organizations such as the American College of Cardiology (ACC) and the European Society of Cardiology (ESC) can play a pivotal role by offering regular workshops, webinars, and courses focused on AI in cardiovascular imaging. These sessions can cover new AI tools, clinical case studies, and interactive discussions on the challenges and benefits of AI integration [72].

Online platforms and resources can also provide accessible training opportunities for healthcare professionals. In addition, developing certification programs for AI proficiency in healthcare can standardize training and ensure a high level of competency among clinicians [73]. Certification can also provide a benchmark for institutions to assess the AI skills of their staff. For instance, a certification program could cover topics such as AI fundamentals, practical applications in cardiovascular imaging, ethical considerations, and patient communication.

Patient involvement and informed consent

As AI technologies, particularly those with black box characteristics, become more integrated into healthcare, especially in cardiovascular imaging, it is crucial to focus on patient involvement and informed consent. Addressing the challenges associated with black box AI models requires specific strategies to ensure patients are informed and engaged in their care decisions [74]. Future efforts should focus on creating and refining communication strategies that help patients understand the use of black box AI in their care. This involves developing educational materials that clearly explain AI technologies, their benefits, risks, and limitations in an accessible manner. Visualization tools, such as interactive diagrams or videos, can be particularly effective in demystifying complex AI concepts [75].

To address the challenges posed by black box AI, the informed consent process must be enhanced. Consent forms should include detailed information about the AI technology being used, how it contributes to the diagnostic or treatment process, and any potential uncertainties or limitations. Future work should explore standardized consent frameworks that can be adapted across various healthcare settings to ensure consistency and thoroughness [76].

Workshops, online courses, and informational brochures can help bridge the knowledge gap and empower patients to participate actively in their care decisions [77]. In addition, future research should aim to make black box AI algorithms more transparent and interpretable to patients. This could involve developing intermediate explanation models or user-friendly interfaces that provide insights into how AI algorithms arrive at their conclusions. For instance, integrating explainable AI (XAI) techniques that generate patient-friendly summaries of the AI's decision-making process can enhance transparency [78].

Standardization, regulatory frameworks and policy aspect in imaging

The lack of standardized criteria for AI explainability presents significant challenges for consistent assessment across various applications [36]. Developing standardized frameworks and metrics for explainability is essential to provide common ground for developers, clinicians, and policymakers [79]. These standards will help ensure that AI models are evaluated consistently, enhancing their reliability and ethical deployment [52].

Ethical challenges posed by unexplainable AI models necessitate robust regulatory frameworks [44]. Future work should focus on creating guidelines that balance transparency, innovation, patient safety, and accountability [66]. Legal frameworks must address the unique complexities of AI in healthcare, ensuring that ethical principles such as beneficence, non-maleficence, autonomy, and justice are upheld in clinical practice. This approach will help build trust among clinicians and patients, facilitating the integration of AI into healthcare workflows [49].

Establishing mechanisms for the regular assessment of AI algorithms will help identify deviations from accepted standards and ensure that AI systems remain aligned with clinical needs [80, 81]. By implementing standardized explainability, robust ethical frameworks, and continuous monitoring, the medical community can ensure the responsible and effective use of AI in cardiovascular imaging and beyond.

Teaching Points:

- Different XAI techniques like model-based and post hoc explanations offer various trade-offs between accuracy and interpretability.
- Understanding both global (dataset-level) and local (instance-level) explanations helps in interpreting how AI models make decisions across different contexts and individual cases.
- Techniques like GRADCAM and feature extraction make complex AI models more understandable by highlighting important features and decision-making processes.
- Combining interpretable models with black box systems and developing user-friendly interfaces can

enhance both transparency and performance in AI applications.

- Comprehensive training for healthcare professionals on AI concepts and applications is crucial for effective and safe AI integration into clinical practice.
- Ensuring that patients are well-informed and involved in decisions regarding AI-based treatments is vital for ethical and effective healthcare delivery.
- Developing standardized criteria and robust regulatory frameworks for AI explainability will help balance innovation with patient safety and ethical considerations.

Conclusion

The integration of AI in cardiovascular imaging holds great potential but is hindered by the black box nature of most of the conventional models used in AI, which poses significant challenges for clinical decision-making, interpretability, and trust. While AI has demonstrated promising results in detecting various cardiovascular conditions, the lack of transparency raises concerns about its reliability and application in evidence-based medicine. To overcome these challenges, there is a pressing need to develop explainable AI (XAI) techniques that provide clear insights into AI decision-making processes. These techniques, including model-based and post hoc explanations, can bridge the gap between complex AI models and the need for transparency in clinical settings.

Moreover, comprehensive education and training programs for healthcare professionals are essential to ensure the effective and responsible use of AI in practice. These programs should equip clinicians with the knowledge and skills to understand and apply AI tools while addressing the ethical implications of their use. Additionally, patient involvement and informed consent must be prioritized to maintain autonomy and trust in AI-driven healthcare.

Finally, establishing robust ethical and regulatory frameworks is crucial for the safe and effective integration of AI in clinical workflows. By addressing these challenges, we can ensure that AI technologies are deployed responsibly, ultimately enhancing patient outcomes and transforming cardiovascular care.

Acknowledgements
Not applicable.

Author contributions

AM: Wrote the introduction, "Bridging the Gap and Future Directions" sections, and the abstract. Created the tables and educational objectives. Reviewed and edited the entire manuscript. PA: Contributed to the conceptualization and design of the study. Conducted the literature search, wrote the initial draft, and reviewed the entire manuscript. ADSA: Contributed to writing the "Black Box Challenges" section of the manuscript. MJE.: Contributed to the conceptualization and design of the study, and co-wrote the "Legal and Ethical Implications" section. AMS: Wrote the "Overview of AI" and "AI Applications in Cardiovascular Imaging" sections. NS: Created the illustrations and reviewed the entire manuscript. MU: Supervised the entire project, suggested the research question, and oversaw the conceptualization, literature search, and manuscript writing. Contributed to writing portions of the manuscript and reviewed and edited the entire manuscript.

Funding
The authors received no financial support for the research, authorship, and/or publication of this article.

Availability of data and material

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details
[1]Faculty of Medicine, Alexandria University, Alexandria, Egypt. [2]Mashhad University of Medical Sciences, Razavi Khorasan, Mashhad, Iran. [3]Faculty of Medicine, Aleppo University of Medical Sciences, Aleppo, Syria. [4]School of Medicine, Isfahan University of Medical Sciences, Esfahan, Iran. [5]Krieger School of Arts and Sciences, Johns Hopkins University, Baltimore, MD, USA. [6]Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA. [7]Russell H. Morgan Department of Radiology and Radiological Sciences, The Johns Hopkins Hospital, Baltimore, MD, USA.

Received: 2 July 2024 Accepted: 1 September 2024
Published online: 13 September 2024

References

1. Can we open the black box of AI?: Nature news & comment, [cited 2024 May 14], Available from: https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731
2. Holzinger A, Biemann C, Pattichis CS, Kell DB (2017) What do we need to build explainable AI systems for the medical domain? [cited 2024 May 14], Available from: https://arxiv.org/abs/1712.09923v1
3. Antoniades C, Oikonomou EK (2024) Artificial intelligence in cardiovascular imaging—principles, expectations, and limitations. Eur Heart J 45(15):1322–6. https://doi.org/10.1093/eurheartj/ehab678
4. Lang M, Bernier A, Knoppers BM (2022) Artificial intelligence in cardiovascular imaging: "unexplainable" legal and ethical challenges? Can J Cardiol 38(2):225–33
5. Seetharam K, Kagiyama N, Sengupta PP (2019) Application of mobile health, telemedicine and artificial intelligence to echocardiography. Echo Res Pract 6(2):R41-52
6. Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP (2018) Machine learning in cardiovascular medicine: are we there yet? Heart 104(14):1156–64
7. Sidey-Gibbons JAM, Sidey-Gibbons CJ (2019) Machine learning in medicine: a practical introduction. BMC Med Res Methodol 19:1–18
8. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M et al (2018) Artificial intelligence in cardiology. J Am Coll Cardiol 71(23):2668–2679
9. Krittanawong C, Johnson KW, Rosenson RS, Wang Z, Aydar M, Baber U et al (2019) Deep learning for cardiovascular medicine: a practical primer. Eur Heart J 40(25):2058–2073
10. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 25.
11. O'Shea K, Nash R (2015) An introduction to convolutional neural networks. arXiv. Available from: http://arxiv.org/abs/1511.08458

12. Schmidt RM (2019) Recurrent neural networks (RNNs): a gentle introduction and overview. arXiv, Available from: http://arxiv.org/abs/1912.05911
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need [Internet]. arXiv; 2023. Available from: http://arxiv.org/abs/1706.03762
14. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S et al. (2014) Generative adversarial networks. arXiv, Available from: http://arxiv.org/abs/1406.2661
15. Danilov VV, Klyshnikov KY, Gerget OM, Kutikhin AG, Ganyukov VI, Frangi AF et al (2021) Real-time coronary artery stenosis detection based on modern neural networks. Sci Rep 11(1):7582
16. Matsumoto T, Kodera S, Shinohara H, Ieki H, Yamaguchi T, Higashikuni Y et al (2020) Diagnosing heart failure from chest X-ray images using deep learning. Int Heart J 61(4):781–786
17. Dikici E, Bigelow M, Prevedello LM, White RD, Erdal BS (2020) Integrating AI into radiology workflow: levels of research, production, and feedback maturity. J Med Imaging 7(1):16502
18. Lin A, Kolossváry M, Motwani M, Išgum I, Maurovich-Horvat P, Slomka PJ et al (2021) Artificial intelligence in cardiovascular imaging for risk stratification in coronary artery disease. Radiol Cardiothorac Imaging 3(1):e200512. https://doi.org/10.1148/ryct.2021200512
19. Marey A, Christopher Serdysnki K, Killeen BD, Unberath M, Umair M, Morgan RH (2024) Applications and implementation of generative artificial intelligence in cardiovascular imaging with a focus on ethical and legal considerations: what cardiovascular imagers need to know! BJR|Artificial Intell. https://doi.org/10.1093/bjrai/ubae008
20. Ghodrati V, Bydder M, Ali F, Gao C, Prosper A, Nguyen KL et al (2021) Retrospective respiratory motion correction in cardiac cine MRI reconstruction using adversarial autoencoder and unsupervised learning. NMR Biomed 34(2):e4433
21. Oscanoa JA, Middione MJ, Alkan C, Yurt M, Loecher M, Vasanawala SS et al (2023) Deep learning-based reconstruction for cardiac MRI: a review. Bioengineering 10(3):334
22. Itu L, Rapaka S, Passerini T, Georgescu B, Schwemmer C, Schoebinger M et al (2016) A machine-learning approach for computation of fractional flow reserve from coronary computed tomography. J Appl Physiol 121(1):42–52
23. Kiryati N, Landau Y (2021) Dataset growth in medical image analysis research. J Imaging 7(8):155
24. Strubell E, Ganesh A, McCallum A (2019) Energy and policy considerations for deep learning in NLP
25. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K (2019) The practical implementation of artificial intelligence technologies in medicine. Nat Med 25:30–6
26. Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25:44–56
27. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A (2019) What clinicians want: contextualizing explainable machine learning for clinical end use. Available from: http://arxiv.org/abs/1905.05134
28. Gallée L, Kniesel H, Ropinski T, Götz M (2022) Artificial intelligence in radiology - Beyond the black box. RoFo Fortschritte auf dem Gebiet der Rontgenstrahlen und der Bildgebenden Verfahren. 195:797–803
29. Götz M, Maier-Hein KH (2020) Optimal statistical incorporation of independent feature stability information into radiomics studies. Sci Rep. https://doi.org/10.1038/s41598-020-57739-8
30. London AJ (2019) Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent Rep 49(1):15–21
31. Nasief H, Zheng C, Schott D, Hall W, Tsai S, Erickson B et al (2019) A machine learning based delta-radiomics process for early prediction of treatment response of pancreatic cancer. NPJ Precis Oncol. https://doi.org/10.1038/s41698-019-0096-z
32. Krittanawong C, Johnson KW, Rosenson RS, Wang Z, Aydar M, Baber U et al (2019) Deep learning for cardiovascularmedicine: a practical primer. Eur Heart J 40:2058-2069C
33. Salih A, Boscolo Galazzo I, Gkontra P, Lee AM, Lekadir K, Raisi-Estabragh Z et al (2023) Explainable artificial intelligence and cardiac imaging: Toward more interpretable models. Circ Cardiovasc Imaging 16(4):E014519
34. Pesapane F, Volonté C, Codari M, Sardanelli F (2018) Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. Insights Imaging 9:745–53
35. Preece A, Harborne D, Braines D, Tomsett R, Chakraborty S (2018) Stakeholders in Explainable AI. 2018. Available from: http://arxiv.org/abs/1810.00184
36. Luo G, Dong S, Wang K, Zuo W, Cao S, Zhang H (2018) Multi-views fusion CNN for left ventricular volumes estimation on cardiac MR images. IEEE Trans Biomed Eng 65(9):1924–1934
37. Nicholson W, Ii P (2017) Artificial intelligence in health care: applications and legal implications. Available from: https://repository.law.umich.edu/articles/1932. Follow this and additional works at: https://repository.law.umich.edu/articles
38. Lang M, Bernier A, Knoppers BM (2022) Artificial intelligence in cardiovascular imaging: "unexplainable" legal and ethical challenges? Can J Cardiol 38:225–33
39. Afifi M, Brown MS (2019) What else can fool deep learning? Addressing color constancy errors on deep neural network performance. Available from: http://arxiv.org/abs/1912.06960
40. Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller KR (2021) Explaining deep neural networks and beyond: a review of methods and applications. Proc IEEE 109(3):247–278
41. Shin M, Kim J, Kim M (2020) Measuring human adaptation to ai in decision making: application to evaluate changes after AlphaGo. Available from: http://arxiv.org/abs/2012.15035
42. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206–15
43. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B (2019) Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci 116(44):22071–22080
44. Sermesant M, Delingette H, Cochet H, Jaïs P, Ayache N (2021) Applications of artificial intelligence in cardiovascular imaging. Nat Rev Cardiol 18:600–9
45. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A et al (2020) Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fus 1(58):82–115
46. Slomka PJ, Miller RJ, Isgum I, Dey D (2020) Application and translation of artificial intelligence to cardiovascular imaging in nuclear medicine and noncontrast CT. Semin Nuclear Med 50:357–66
47. Hagen GR (2021) AI and Patents and Trade Secrets. In: Martin-Bariteau Florian, Scassa Teresa, editors. Artificial Intelligence and the Law in Canada. Toronto
48. Reznick RK, Harris K, Horsley T (2020) Task force report on artificial intelligence and emerging digital technologies
49. Cestonaro C, Delicati A, Marcante B, Caenazzo L, Tozzo P (2023) Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review. Front Med 10:1305756
50. Zheng Q, Delingette H, Ayache N (2019) Explainable cardiac pathology classification on cine MRI with motion characterization by semi-supervised learning of apparent flow. Med Image Anal 1(56):80–95
51. Zhang Q, Hann E, Werys K, Wu C, Popescu I, Lukaschuk E et al (2020) Deep learning with attention supervision for automated motion artefact detection in quality control of cardiac T1-mapping. Artif Intell Med 110:101955
52. Geis JR, Brady AP, Wu CC, Spencer J, Ranschaert E, Jaremko JL et al (2019) Ethics of artificial intelligence in radiology: summary of the joint European and north American multisociety statement. J Am Coll Radiol 16(11):1516–1521
53. Schoepf UJ. Contemporary Medical Imaging Series Editor. Available from: https://link.springer.com/bookseries/7687
54. Reznick RK, Harris K, Horsley T (2020) Task force report on artificial intelligence and emerging digital technologies
55. Khoury Lara (2006) Uncertain causation in medical liability. Hart Pub. p 270
56. Frank X (2019) Is Watson for oncology per se unreasonably dangerous?: Making a case for how to prove products liability based on a flawed artificial intelligence design. Am J Law Med 45(2–3):273–294
57. Van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA (2022) Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal 79:102470
58. Lage I, Chen E, He J, Narayanan M, Kim B, Gershman SJ et al. (2019) Human evaluation of models built for interpretability. In: Proceedings of

the AAAI conference on human computation and crowdsourcing. pp 59–67

59. Abbasi-Asl R, Yu B (2017) Structural compression of convolutional neural networks. arXiv preprint arXiv:170507356
60. Olden JD, Joy MK, Death RG (2004) An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol Modell 178(3–4):389–397
61. Petsiuk V, Jain R, Manjunatha V, Morariu VI, Mehra A, Ordonez V et al. (2021) Black-box explanation of object detectors via saliency maps. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 11443–52.
62. Tsang M, Cheng D, Liu Y (2017) Detecting statistical interactions from neural network weights. arXiv preprint arXiv:170504977.
63. Olah C, Mordvintsev A, Schubert L (2017) Feature visualization. Distill 2(11):e7
64. Clough JR, Oksuz I, Puyol-Antón E, Ruijsink B, King AP, Schnabel JA (2019) Global and local interpretability for cardiac MRI classification. In: International conference on medical image computing and computer-assisted intervention. Springer. pp 656–64.
65. Kingma DP, Welling M (2019) An Introduction to variational autoencoders. Found Trends Mach Learn 12(4):307–92
66. Hybrid decision making: when interpretable models collaborate with black-box models | DeepAI. [cited 2024 Jun 5]. Available from: https://deepai.org/publication/hybrid-decision-making-when-interpretable-models-collaborate-with-black-box-models
67. Gadzinski G, Castello A (2022) Combining white box models, black box machines and humaninterventions for interpretable decision strategies. Judgm Decis Mak 17(3):598–627
68. Charow R, Jeyakumar T, Younus S, Dolatabadi E, Salhia M, Al-Mouaswas D et al (2021) Artificial intelligence education programs for health care professionals: scoping review. JMIR Med Educ 7(4):e31043
69. Amann J, Blasimme A, Vayena E, Frey D, Madai VI (2020) Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak. https://doi.org/10.1186/s12911-020-01332-6
70. van Kooten MJ, Tan CO, Hofmeijer EIS, van Ooijen PMA, Noordzij W, Lamers MJ et al (2024) A framework to integrate artificial intelligence training into radiology residency programs: preparing the future radiologist. Insights Imaging 15(1):1–14. https://doi.org/10.1186/s13244-023-01595-3
71. RSNAI | RSNA. [cited 2024 Jun 5]. Available from: https://www.rsna.org/rsnai
72. Training and Education: Provide training and education for healthcare professionals to understand and trust AI applications. [cited 2024 Jun 5]. Available from: https://www.researchgate.net/publication/378342227_Training_and_Education_Provide_training_and_education_for_healthcare_professionals_to_understand_and_trust_AI_applications
73. De Cecco CN, van Assen M, Leiner T, editors (2022) Artificial Intelligence in Cardiothoracic Imaging. [cited 2024 Jun 5]. https://doi.org/10.1007/978-3-030-92087-6
74. Fenech ME, Buston O (2020) AI in cardiac imaging: A UK-based perspective on addressing the ethical, social, and political challenges. Front Cardiovasc Med 15(7):508920
75. Zhang W, Cai M, Lee HJ, Evans R, Zhu C, Ming C (2024) AI in Medical Education: Global situation, effects and challenges. Educ Inf Technol 29(4):4611–33. https://doi.org/10.1007/s10639-023-12009-8
76. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN et al (2023) Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 23:1–15. https://doi.org/10.1186/s12909-023-04698-z
77. Chan B (2023) Black-box assisted medical decisions: AI power versus ethical physician care. Med Health Care Philos 26(3):285–92. https://doi.org/10.1007/s11019-023-10153-z
78. Frasca M, La Torre D, Pravettoni G, Cutica I (2024) Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review. Discover Artif Intell 4(1):1–21. https://doi.org/10.1007/s44163-024-00114-7
79. Pesapane F, Volonté C, Codari M, Sardanelli F (2018) Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. Insights Imaging 9:745–53
80. Reznick RK, Harris K, Horsley T. Task force report on artificial intelligence and emerging digital technologies
81. Al'Aref SJ, Anchouche K, Singh G, Slomka PJ, Kolli KK, Kumar A et al (2019) Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. Eur Heart J 40:1975–86

Publisher's Note