IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 January 2025, accepted 19 March 2025, date of publication 27 March 2025, date of current version 8 April 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3555543

SURVEY

State-of-the-Art in Responsible, Explainable, and Fair AI for Medical Image Analysis

SOHEYLA AMIRIAN[1], FENGYI GAO[2], NICKOLAS LITTLEFIELD[3,4], (Graduate Student Member, IEEE), JONATHAN H. HILL[1], ADOLPH J. YATES JR.[5], JOHANNES F. PLATE[5], LIRON PANTANOWITZ[4,6], HOOMAN H. RASHIDI[4,6], AND AHMAD P. TAFTI[2,3,4,6]

[1]Seidenberg School of Computer Science and Information Systems, Pace University, New York, NY 10038, USA
[2]Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA 15260, USA
[3]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15260, USA
[4]Computational Pathology and AI Center of Excellence (CPACE), University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA
[5]Department of Orthopaedic Surgery, University of Pittsburgh, Pittsburgh, PA 15213, USA
[6]Department of Pathology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA

Corresponding authors: Soheyla Amirian (samirian@pace.edu) and Ahmad P. Tafti (tafti.ahmad@pitt.edu)

ABSTRACT Integrating responsible, explainable, and fair artificial intelligence (REF-AI) into medical image analysis has gained significant attention in recent years. This has been driven by the pressing need for ethical, trustworthy, and transparent implementation of AI systems in healthcare. The following review provides a concise overview of REF-AI in the context of medical image analysis. It begins with the fundamental concepts of AI responsibility, explainability, and fairness, followed by a comprehensive taxonomy of over 35 key algorithms and strategies. In addition, it compares methodologies, strengths, and limitations, such as the alignment of AI models with medical standards and the development of interpretable and actionable results for clinicians. Finally, it highlights current trends and proposes directions for future research to further advance the responsible, explainable, and fair application of AI in medical imaging.

INDEX TERMS Artificial intelligence (AI), responsible AI, explainable AI, fair AI, AI evaluation metrics, medical imaging.

I. INTRODUCTION

Integrating artificial intelligence (AI) into healthcare, particularly in medical image analysis, has made transformative advances in diagnosis, prognosis, treatment planning, and patient care. As these advanced technologies continue to evolve, there is a pressing need for AI systems that are effective, ethical, transparent, trustworthy, and equitable. Responsible, explainable, and fair AI (REF-AI) has emerged as a critical focus in ensuring that AI applications in medical imaging meet these standards. REF-AI emphasizes the importance of aligning AI models with well-established medical standards, making their outputs interpretable and actionable for healthcare professionals, and ensuring fairness in their application across diverse patient populations and healthcare settings. This review explores the foundational principles of REF-AI in medical image analysis, providing a comprehensive overview of the key algorithms and methodologies that drive this field. This review aims to advance the responsible, explainable, and equitable application of AI in medical imaging by critically assessing the strengths, limitations, and challenges of existing approaches while proposing directions for future research. To the best of our knowledge, our research is the first to focus specifically on REF-AI within the domain of medical image analysis. We will begin by defining REF-AI and exploring its empirical foundations.

A. RESPONSIBLE AI IN MEDICAL IMAGE ANALYSIS

Responsible AI refers to building, implementing, and utilizing AI systems in an ethical, sustainable, and trustworthy manner. This focuses on respecting human values, promoting

The associate editor coordinating the review of this manuscript and approving it for publication was Fan-Hsun Tseng.

fairness, and ensuring transparency and accountability while considering societal impacts [1], [2], [3], [4]. Responsible AI in healthcare encompasses the broader ethical and social responsibilities of developing and deploying AI-powered systems across the healthcare ecosystem. This empowers individuals to comprehend, control, and take responsibility for AI-driven mechanisms [1], [5], [6]. Responsible AI in medical imaging ensures that AI tools are implemented and used in ways that align with medical ethics, fairness and equity, regulations, patient privacy, and safety standards [7], [8], [9], [10]. This includes comprehensive and even rigorous testing, validation, and continuous monitoring of AI systems to guarantee that they perform precisely, logically, and safely in real-world clinical settings.

B. EXPLAINABLE AI IN MEDICAL IMAGE ANALYSIS

AI explainability in healthcare refers to AI models' capability to provide clear and understandable explanations of their decisions and predictions to their end-users, such as clinicians, surgeons, and patients [11], [12], [13], [14]. In AI-powered image analysis [15], [16], [17], [18], and particularly for medical image analysis, explainable AI is critical because healthcare professionals must trust and understand the AI-powered imaging toolsets before they utilize them for patient care and clinical practices. This could bridge the gap between complex AI models, domain experts, clinicians, patients, and decision-makers by providing insights into how and why a particular diagnosis or prediction was made using the AI models [14], [19], [20], [21], [22].

C. FAIR AI IN MEDICAL IMAGE ANALYSIS

AI fairness involves addressing biases in data, data acquisition (e.g., imaging machinery), algorithms, and outcomes to prevent discrimination [23], [24], [25]. In medical image analysis, this involves training and validating AI models on diverse datasets that reflect various demographics, such as sex, race, and ethnicity, alongside social determinants of health (SDOH), such as income and social protection, education, availability of healthcare services, health insurance, and quality of care [26], [27]. This method helps reduce the risk of biased predictions that might unequally impact specific patient groups. Fairness in medical imaging is essential to verify that AI tools provide equal benefits to all patients, regardless of their race, sex, age, economic status, or other factors [28], [29], [30], [31].

These three principles of responsibility, explainability, and fairness confirm that AI systems in medical imaging are technically and computationally effective, ethically sound, and well-equipped with a list of qualitative and quantitative attributes. These attributes include accountability, lawfulness, traceability, reliability, equity, governability, scalability, availability, explainability, truthfulness, privacy-preserving, and safety (Figure 1). These are essential to gain acceptance among healthcare professionals and patients and seamlessly integrate AI into various clinical workflows.

FIGURE 1. Principles for REF-AI in medical imaging include accountability, privacy-preserving, lawfulness, traceability, reliability, equity, governability, scalability, availability, explainability, truthfulness, and safety.

D. EMPIRICAL FOUNDATIONS

The empirical foundation of responsible, explainable, and fair AI (REF-AI) in medical image analysis lies in creating AI models and mechanisms that are reliable, understandable, and fair in various healthcare settings and patient populations. With that, responsible AI emphasizes safety and ethical alignment by evaluating AI tools against real-world clinical standards to ensure accuracy and reliability. Regular monitoring confirms that AI outputs remain dependable as new patient data or clinical scenarios emerge. Meanwhile, explainable AI makes AI decisions understandable and accessible to healthcare providers and patients, enabling them to comprehend how specific features led to a particular result. This transparency confidently supports clinicians in integrating AI insights into their decision-making. Fair AI, on the other hand, ensures that these systems perform equitably for all patients. Fairness reduces the risk of bias by training AI on diverse datasets representing different ages, ethnicities, and health conditions. It brings consistent performance in all groups, promoting equitable patient care.

Specific metrics and measurements should be used to evaluate each component to assess the effectiveness of REF AI in medical imaging. For AI responsibility, metrics, such as accuracy, sensitivity, and specificity, determine if an AI model delivers reliable, accurate, and clinically relevant output. These metrics ensure that AI predictions align with accepted medical standards and provide consistent support in diverse clinical scenarios. Regarding AI explainability, some computational strategies, such as saliency maps [32], [33] and Grad-CAM [34], [35], help clinicians visualize the AI's decision-making process by highlighting the specific

areas of a medical image that contributed to a diagnosis. It offers better transparency and allows healthcare providers to interpret AI recommendations relatively easily and quickly. Evaluation tools, such as the Pointing Game [36], [37] or Area Over Perturbation Curve (AOPC) [38], measure how well these visual explanations align with medical insights. The fairness of AI could also be evaluated using fairness-adjusted metrics. For example, using skew equalized odds metric [39], we can compare AI performance across different demographic groups to ensure equitable treatment recommendations. Additionally, skew error ratio (SER) [28] can measure the discrepancy in error rates between different demographic groups, such as age, gender, or ethnicity, helping to identify potential biases in the AI model's performance. Moreover, metrics like demographic parity and subgroup accuracy help identify potential bias, ensuring that the AI models serve all patient populations effectively and without discrimination.

E. REF-AI APPLICATIONS IN MEDICAL IMAGING

By making AI technologies accurate, efficient, and in line with clinical requirements and regulatory frameworks, REF-AI is now transforming medical imaging into more reliable, more intelligent, and more responsible pipelines in various clinical domains. With reliable predictions, transparent reasoning, and equitable performance, REF-AI provides a basis for safe and practical applications in diagnosis, prognosis, personalized treatment, and imaging, fundamentally enhancing patient care and clinical outcomes.

This section organizes REF-AI applications for medical imaging into five categories, as follows:

- **Diagnosis, Prognosis, and Treatment:** REF-AI supports accurate and equitable diagnostic tools, particularly identifying and predicting disease progression. For instance, REF-AI in diagnostic imaging provides clinicians with clarity by visually highlighting regions of interest (RoI) in medical images, such as cancerous lung nodules, lesions, or abnormal tissue, allowing for faster and more informed clinical decisions. REF-AI techniques improve prognostic predictions while estimating disease progression with high accuracy, which aligns with real-world medical standards. It also makes it possible for clinicians and surgeons to choose treatment paths that reflect patient-specific needs, reducing biases that could adversely impact diagnosis and prognosis across diverse patient demographics [19], [31], [40], [41], [42], [43], [44], [45].
- **Personalized Medicine:** Personalized care is becoming a reality with REF-AI, as AI systems harness imaging data combined with other patient characteristics, including genetics and medical history, to custom treatment strategies. For example, REF-AI methods clarify the reasoning behind these suggestions, enabling clinicians and patients to understand the factors driving personalized recommendations. These AI-enabled systems make individualized care models available and equally beneficial across all demographic groups, optimizing patient outcomes without compromising equity in care [44], [45], [46], [47], [48], [49], [50].
- **AI Assistive Radiology:** REF-AI significantly improves radiology by offering AI-based assistance that radiologists can trust and use intuitively in everyday practice. These technologically advanced tools highlight critical areas within radiological images, such as suspicious masses and bone or prosthesis loosening, enhancing the radiologist's ability to make fast and accurate diagnoses. These AI models are validated against diverse clinical imaging datasets to produce consistent and high-quality support, while they are equally accurate and applicable across various patient populations [42], [50], [51], [52], [53], [54], [55], [56], [57].
- **Medical Image Segmentation and Classification:** REF-AI has made a significant leap towards more actionable and interpretable outcomes in segmentation and classification. REF-AI tools are now helping radiologists and other clinicians understand how AI models segment, measure, and classify regions, such as highlighting tumor boundaries in cancer imaging, which aids in diagnosis and treatment planning. These computational methods focus on reliable and real-world testing, aligning the automatic segmentation algorithms closely with clinical benchmarks and patient-specific characteristics. By reducing biases and making generalized AI models, REF-AI advances image segmentation and classification accuracy across diverse populations, eliminating variability in AI model performance that could otherwise lead to mis-classification or even inconsistent patient outcomes [28], [42], [43], [44], [51], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68].
- **Clinical Trials:** REF-AI plays an important in clinical research by supporting unbiased patient selection and reliable outcome prediction, both essential for generalizable and reproducible results. For instance, REF-AI mechanisms assess patient data in a way that reduces demographic or socioeconomic biases, making trials more inclusive and representative. At the same time, these AI-powered methods offer transparency into AI-driven insights built from imaging data, helping researchers interpret these insights and strengthen the credibility of trial outcomes. This application reduces trial dropout rates and enables more precise tracking of intervention effects, advancing research quality and paving the way for broader application in clinical practice [23], [24], [69], [70], [71], [72], [73], [74].

Our key contributions are summarized as follows.

- **Comprehensive Overview of REF-AI in Medical Imaging:** This work provides an in-depth review of the foundational principles of Responsible, Explainable, and Fair AI (REF-AI) in medical image analysis.
- **Analysis of Key Algorithms and Methodologies:** We examine the core computational techniques that advance

REF-AI applications in medical imaging, highlighting their strength and limitations.

- **Advancing Ethical AI Implementation in Healthcare:** We highlight the significance of ethical AI practices in healthcare, advocating for responsibility, accountability, transparency, and fairness in AI applications.
- **Comprehensive Taxonomy of REF-AI Algorithms:** We introduce a detailed taxonomy encompassing over 35 key algorithms and strategies, categorizing them based on their role in building responsibility, explainability, and fairness in AI-driven medical imaging. See Figure 2.
- **Future Research Directions:** We propose pathways for enhancing REF-AI, including novel frameworks, interdisciplinary collaborations, and policy recommendations.

II. REF-AI CHARACTERISTICS

This section further explains the underlying components of REF-AI in medical imaging. This section will offer a comprehensive explanation of the REF-AI principles depicted in Figure 1.

A. ACCOUNTABILITY

Accountability in REF-AI refers to AI systems that can be audited to meet specific standards, with clearly defined responsibilities and consequences if the systems fail to comply. It also discusses that AI developers, organizations, and policymakers are legally and ethically responsible for the decisions made by AI systems, requiring them to follow laws and standards to ensure AI-powered systems function correctly [75], [76], [77]. The establishment of REF-AI systems should hold developers and organizations accountable for errors or issues, with accountability mandated at every level from requirement analysis and development to implementation and deployment [78]. There is also a pressing need to organize a solid data governance framework with regular data audits to assess data records' quantity, quality, and suitability, equipping such a framework with data provenance, data traceability, and a comprehensive data dictionary and documentation [79]. One way to achieve accountability in AI systems for medical imaging is by ensuring AI decisions are transparent and explainable to end-users (e.g., physicians, radiologists, and clinicians). This allows them to review, investigate, and validate the AI's conclusions before conveying information to patients. This collaborative process adds an important layer of human oversight, confirming that AI tool sets are used responsibly, and they can deliver accurate and trustworthy medical guidance [80], [81].

B. PRIVACY-PRESERVING

Privacy-preserving AI in REF-AI refers to developing and deploying AI systems that safeguard sensitive patient information while maintaining the system's ability to deliver reliable, trustworthy, accurate, and actionable insights. Given the highly confidential nature of healthcare data, including medical images combined with clinical notes or radiology reports, ensuring privacy preservation is essential for compliance with legal and ethical standards, such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR). Beyond regulatory compliance, privacy-preserving AI builds trust among stakeholders, including patients, surgeons, clinicians, and healthcare organizations, sharing data with AI algorithms in a way that cannot be traced back to individuals. This trust is fundamental to the successful and widespread integration of AI technologies in medical imaging informatics and healthcare [82], [83], [84]. Practical methods of privacy-preserving AI are rapidly growing in medical imaging. For example, federated learning and homomorphic encryption have been utilized for training AI models across multiple institutions, allowing collaboration while securing patient data [85], [86], [87], [88], [89].

The federated learning approach allows AI models to be trained on distributed data sources without transferring data via a central server. This allows sensitive patient data to remain within the institution that owns it [85] and [86]. Additionally, with computational methods such as encrypted data processing using homomorphic encryption, we can make computations directly on encrypted data, safeguarding privacy throughout the AI pipeline [87], [88]. Research using AI and associated databases will exponentially increase; safeguards regarding compliance with national and international rules and regulations regarding human research need to be maintained.

C. LAWFULNESS

Lawfulness in AI refers to aligning AI systems with established legal and regulatory frameworks, guaranteeing adherence to norms and standards that protect patients and uphold ethical principles. In medical imaging, where AI systems analyze sensitive patient data and assist in clinical decision-making, lawfulness becomes paramount. Compliance with regulations, such as the GDPR in the European Union [90], emphasizes that AI must manage health data records responsibly. The GDPR sets comprehensive rules on data privacy and prohibits automated decisions without human involvement unless exceptions apply, such as explicit patient consent or justified public interest [91], [92]. By preserving patient rights and promoting confidence in AI-assisted healthcare, these legal protections aim to ensure AI-powered medical imaging methods function within an ethical and legal framework.

Furthermore, for AI to be legal, accountability and transparency must be precisely maintained for the decisions these systems make. In medical imaging, for example, they must produce outputs that are understandable to medical practitioners, allowing for significant human oversight of the

AI-enabled decision-making process. When medical professionals can verify AI-driven conclusions, the system's outputs become more credible, allowing for REF-AI integration into diagnostic, prognostic, and therapeutic processes [92], [93]. Moreover, lawful AI systems must account for justice and nondiscrimination to prevent disparities, particularly in diverse healthcare settings and patient populations. Such a commitment to fairness and nondiscrimination is central to lawful AI, as a biased AI algorithm in medical imaging may lead to inconsistent image segmentation, diagnoses, and treatments [28]. Together, these characteristics of lawfulness in AI uphold both the ethical and legal standards essential to patient care, public trust, and positive societal impact [92], [93], [94], [95].

D. TRACEABILITY

Building REF-AI systems, especially in the high-stakes healthcare industry, requires traceability. Traceability in AI involves maintaining a comprehensive record of each phase in the system's life cycle, starting from functional and nonfunctional requirement analysis, system specification, and data collection to AI model development, validation, and deployment [96]. Traceability is an essential safety measure in AI-powered medical imaging, particularly where AI aims at sensitive diagnostic and treatment strategies. It guarantees that any mistakes or discrepancies can be traced to a particular phase of the system's life cycle or an incorrect decision within the AI pipeline, thus promoting adherence to patient safety, clinical standards, and healthcare regulations [97].

To achieve traceability in REF-AI in healthcare, AI systems should be equipped with comprehensive documentation and version control at each stage of development. This includes maintaining detailed records of data sources, core functionalities of the system, training procedures, AI model parameters, and validation strategies [5], [97], [98]. In medical imaging analysis, traceability is particularly critical due to the complex nature of interpreting medical images and the potentially severe consequences of misdiagnosis or misclassification. By tracking each step, from data selection to automatic image analysis and diagnostic outputs, traceability allows clinicians and regulators to identify the source of errors and any biases in AI recommendations, facilitating accountability and ethical practice in patient care [5], [97], [98], [99], [100].

E. RELIABILITY

Reliability in REF-AI systems in healthcare means consistently performing as expected across diverse healthcare settings and scenarios, managing uncertainties and errors, and preventing breakdowns or performance issues to maintain core functionalities [101], [102]. In the context of AI, it demonstrates that a system can repeatedly provide accurate and stable outputs, even when faced with unexpected conditions. This is particularly important in medical image analysis, where AI tool sets assist with diagnostics, treatment planning, and patient monitoring, and where unreliable outcomes could pose significant clinical risks [96], [103], [104], [105].

F. EQUITY

Equity in REF-AI, particularly in healthcare, addresses fair access, representation, and outcomes across diverse populations. Equity in AI involves actively mitigating biases that arise from unequal data representation or algorithmic bias. Implementing AI with a foundation of equity is a key component in REF-AI, where studies emphasize that AI systems developed without sufficient regulatory oversight can inadvertently propagate health inequities, negatively affecting marginalized and historically underrepresented groups [106]. In medical imaging, equity is significant as AI-powered models have been used successfully for diagnostics and predictive analysis. Equity-focused AI policies and strategies advocate for regulatory mechanisms requiring diverse data sources and imaging machinery in AI model training. These standards mitigate biases, demonstrating that AI technologies can actively address healthcare disparities [2], [8], [28], [79], [103], [107], [108].

REF-AI offers the potential to advance equity in AI for medical imaging by proposing methods for bias detection and mitigation at every development phase starting from data collection, to AI model development, evaluation, deployment, and interpretation. This includes utilizing federated learning or cyclic weight transfer approaches, allowing institutions to share model knowledge without transferring personal data, thus supporting fair representation while maintaining privacy [107], [109], [110]. By building a regulatory environment prioritizing fairness, REF-AI can act as a revolutionary force in healthcare. This is particularly true in applications such as medical imaging, where accurate, reliable, and de-biased decision-making is important for patient outcomes across all populations within different healthcare systems.

G. GOVERNABILITY

Governability in AI is critical to maintaining accountability, upholding ethical standards, and building trust within healthcare systems. The rapid integration of AI in healthcare and medical imaging analysis begs a robust governance model that addresses ethical and practical challenges [111], [112]. Without clear regulations, AI systems may result in biases that compromise patient privacy and/or diminish clinical confidence. A well-structured and well-organized governance framework is thus needed to build AI applications operating under strict healthcare standards and regulations while being adaptable to emerging challenges. In medical imaging, where accuracy is paramount, governability directly impacts AI-powered diagnostics' reliability, availability, and fairness (e.g., cancerous lung nodules, and bone lucency). Effective governance frameworks reduce risks associated with biased or non-representative images, supporting diagnostic accuracy

across diverse patient groups. By embedding these practices, REF-AI in medical imaging can consistently deliver impartial and interpretable results, enhancing diagnostic accuracy and reinforcing trust between patients, clinicians, and healthcare providers [113]. Moreover, a comprehensive governance model that includes inputs from all parties, such as clinicians, AI developers, and patient representatives, helps align AI applications with ethical healthcare principles. Mechanisms such as regular audits, performance evaluations across different demographic groups (e.g., sex, race, age, ethnicity), and adherence to privacy make a platform to integrate AI into clinical workflows smoothly. This requires careful attention, as an overly rigid framework here can slow down advancements and innovations in AI integration. At the same time, insufficient oversight may risk patient safety and perhaps ethical breaches [114], [115], [116], [117].

H. SCALABILITY

Scalability in AI refers to the capacity of AI systems to efficiently handle growing workloads and user numbers and maintain consistent performance as they expand. For medical imaging analysis, scalability is essential due to the growing volume of imaging data and the computational demands of advanced AI methods, such as deep convolutional neural networks (CNNs). The scalability of AI methods now mainly relies heavily on cloud infrastructure, parallel processing, and optimized data management. This collectively enables the high throughput necessary for fast and real-time medical analysis and diagnostics [118], [119], [120], [121]. Achieving scalable AI in medical imaging involves the adoption of advanced deep learning practices, including transfer learning and federated learning, which allow models to train on varied datasets without centralized data pooling [122], [123], [124], [125]. In healthcare, AI scalability presents challenges as well. For example, the heterogeneity of medical imaging data requires systems to accommodate different image resolutions, diverse image qualities, and labeling conventions. Cloud-based solutions offer flexible resources and facilitate collaboration across institutions, addressing the high storage and processing requirements while also supporting interoperability across healthcare systems [121].

I. AVAILABILITY

REF-AI should come with AI availability, where it mainly aims to build AI tool sets that are easy to use and accessible for all types of healthcare providers, regardless of technology or data system's level, making them available for all parties [126], [127], [128], [129]. Different healthcare providers, ranging from big hospitals to smaller clinics, may have vastly different computer systems, equipment, and data types. For AI to be helpful across the board, it must be designed to work well in all these diverse settings. That means building AI tools that are flexible and can also fit smoothly into any healthcare setting; thus, even less-resourced or resource-limited clinics can still benefit from AI-powered technologies.

AI availability also refers to the fact that AI systems can operate across different platforms, such as various operating systems (e.g., Linux, Windows), imaging machines (e.g., MRI, X-ray), and smart devices. This flexibility is needed to meet the demands of diverse clinical environments where technology and equipment can differ significantly. By making AI accessible across these different devices and systems, healthcare organizations can provide clinicians with consistent and reliable AI-powered support in diagnostics, decision-making, and patient monitoring, regardless of their specific tools or machines. Furthermore, the 24/7 availability of AI solutions is crucial in healthcare, as prompt access to AI-driven insights can have a significant impact on patient outcomes during emergencies. Finally, the availability of AI in healthcare is incomplete without addressing AI literacy and education among healthcare professionals. For AI to be successfully adopted, healthcare providers and staff need to understand how to use and benefit from these advanced technologies effectively. This requires investing in AI literacy programs to equip clinicians, technicians, and administrative staff with the knowledge to utilize AI tools confidently. When healthcare providers across the spectrum are trained in AI applications and understand the underlying principles, the sector can move toward a more informed and proactive approach to adopting and implementing AI innovations.

J. EXPLAINABILITY

Explainability in REF-AI clarifies how AI systems make decisions, process data and images, and function internally. By providing transparent explanations of each decision-making step, REF-AI allows users to see exactly how data is processed, what factors contribute to AI-powered decisions, and how results are generated, making AI components more understandable for the end-users (e.g., clinicians, patients, physicians) [22], [103], [130], [131], [132], [133], [134]. This focus on transparency enables healthcare providers and patients to gain insights into the rationale behind AI-powered decisions and the meaning of these decisions. In this way, REF-AI explainability bridges the gap between complex AI systems and end-users and makes more ethical and informed use of AI, supporting decision-making in high-stakes areas like medical imaging analysis [22], [135], [136], [137], [138], [139].

In medical imaging, REF-AI's explainability feature is especially valuable for surgeons, radiologists, and patients, as it helps them understand AI-generated diagnoses or recommendations. This understanding contributes more to shared decision-making, allowing medical professionals to confidently incorporate AI insights while patients feel more informed about their care. By explaining how the AI arrived at each conclusion, REF-AI also reduces the risk of errors or misdiagnosis, while it can also make these AI-enabled models transparent and free from biases [14], [19], [28], [140], [141], [142], [143], [144], [145].

K. TRUTHFULNESS

Truthfulness in REF-AI means the ability of AI systems to provide accurate, consistent, and fact-based outputs aligned with medical knowledge and clinical evidence. Medical imaging helps to build AI-powered medical imaging algorithms that are correct and free from misleading or even exaggerated results. Without truthfulness, AI systems risk disseminating inaccuracies that could lead to diagnostic errors (e.g., misclassification), ineffective treatments, prolonged hospitalization, and reduced patient confidence in AI-powered healthcare [66], [140], [146], [147]. AI truthfulness in medical imaging can be illustrated through practical examples. For instance, AI systems that detect cancerous lung nodules from CT images must generate predictions based on verified diagnostic markers, such as size, morphology, texture, or masses, while avoiding false positives caused by image artifacts [148], [149]. Another example is the use of AI for brain tumor segmentation and classification, where truthfulness ensures that highlighted tumor regions correspond accurately to the pathology confirmed through biopsy or clinical evaluation [150], [151], [152].

Implementing truthfulness in AI systems for medical imaging informatics should include several key elements, including but not limited to (1) rigorous training and validation of AI models using diverse and high-quality datasets to confirm that predictions are based on reliable data representative of real-world scenarios, (2) adherence to medical guidelines and standards, such as clinical protocols, thus it can align AI outputs with currently accepted truths in healthcare, (3) robust error analysis, validation mechanisms, and continuous monitoring for AI-enabled methods, (4) integrating strategies for peer review and clinical oversight to keep AI systems accountable to healthcare providers [147], [153], [154], [155], and (5) ensure consensus processes that allow incorporation of new knowledge (some generated by widespread use of AI) with consistency across platforms.

L. SAFETY

Safety in REF-AI involves building and implementing AI systems that are robust, dependable, and free from behaviors or predictions that could lead to harm. In medical imaging, where AI supports critical diagnostic and treatment decisions, safety is of paramount importance [111], [156], [157], [158]. Unsafe AI outputs, such as inaccurate predictions, incorrect tumor segmentation, or misclassification of disease, could result in delayed treatments, misdiagnoses, or costly and unnecessary procedures. We can protect patients and uphold clinical standards by embedding safety principles into AI systems, enhancing trust among healthcare providers and stakeholders. In medical imaging informatics, for example, AI models for lung nodule detection on CT scans may employ confidence thresholds to identify and triage cases requiring additional human review. Additionally, it can prompt clinicians to cross-reference with other imaging modalities or clinical data, thus reducing the risk of false negative classification.

Regarding safety implementation, robust model validation techniques and testing against diverse and extensive datasets will help consistent performance across various patient demographics and imaging scenarios. Moreover, employing adversarial robustness techniques, such as adversarial training or quantifying errors in predictions, can protect AI systems from malicious inputs that could compromise their accuracy. Continuous monitoring, regular updates, and retraining of AI models on new data can also help maintain safety as clinical practices and/or patient data evolve.

III. REF-AI METHODOLOGIES

This section dissects the methodologies associated with REF-AI frameworks and approaches. Through our analysis of these various methodologies, we aim to emphasize best practices and emerging trends that support effective AI technologies. This review offers a foundational understanding of how these principles can be incorporated into AI systems, enhancing trust and accountability in their applications. Figure 2 illustrates the taxonomy of REF-AI methodologies for medical imaging informatics. Table 1 provides an overview of the key methodologies, strategies, and techniques categorized under the REF-AI framework. The table serves as a concise reference for understanding how REF-AI methodologies contribute to developing equitable, interpretable, and trustworthy AI systems in medical imaging informatics.

A. RESPONSIBLE AI

Responsible AI methods encompass various strategies, including bias detection and mitigation, transparency tools, data privacy, and ethical frameworks, designed to ensure that AI systems are developed and deployed, emphasizing fairness, transparency, and ethical standards, especially in medical imaging. This section discusses methodologies designed to ensure AI responsibility.

1) BIAS DETECTION AND MITIGATION

Bias mitigation methods are strategies employed first to identify and reduce AI system biases. These biases can arise from data-driven algorithmic and human biases [95], [159]. Ensuring the diversity among patients, stakeholders, healthcare workers, and the datasets used to train and validate machine learning models can help reduce biases and increase equity in outcomes [160], [161]. A list of mechanisms exists to cope with bias detection and mitigation. For instance, causal models and graphs can detect and address direct discrimination in data, revealing hidden biases and enabling corrective actions [162]. Generating synthetic data helps supplement real-world data, particularly when underrepresented groups lack representation. Resampling

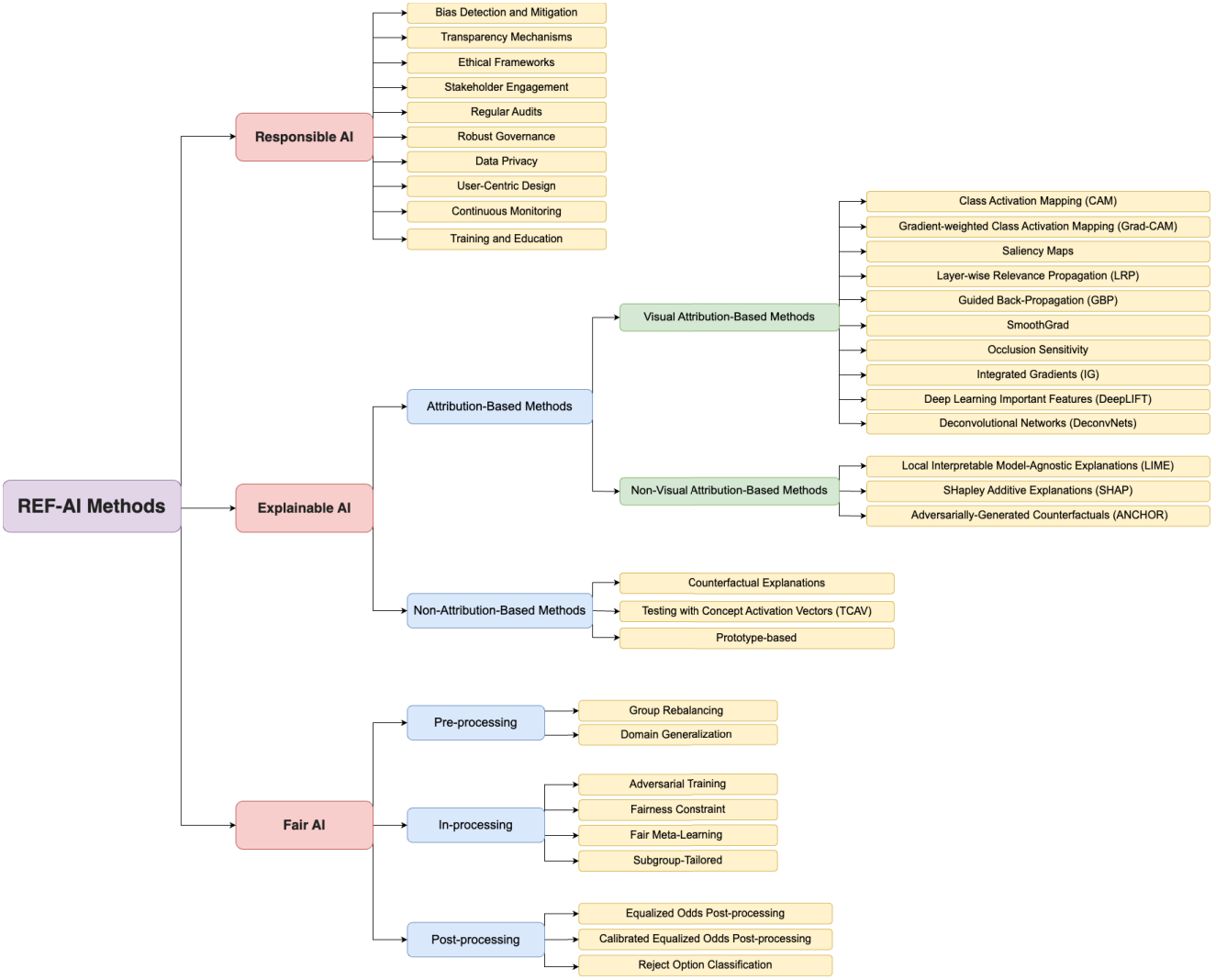


FIGURE 2. The taxonomy of the responsible, explainable, and fair artificial intelligence (REF-AI) methods for medical image analysis.

techniques, including down-sampling and oversampling, can address imbalances in the training data [163]. Furthermore, careful handling of missing data, particularly for marginalized populations, is essential to avoid introducing further bias, acknowledging that missing data is often non-random and requires thoughtful treatment [79], [162]. From an algorithm perspective, algorithmic adjustments, such as IBM's ''adversarial debiasing'' technique, can act as an adversary by predicting sensitive attributes like race or sex/gender from the data and mitigating their influence on any predictions [161].

Developing solid guidelines emphasizing data origin and quality can help organizations select high-quality datasets. When domain experts and annotators create gold-standard or ground-truth datasets, understanding their experiences and measuring inter-rater agreement becomes important for maintaining data quality and minimizing biases [28], [79], [162].

2) TRANSPARENCY MECHANISMS

Transparency mechanisms are essential in responsible AI to enhance the interpretability and accountability of AI models. Transparency mechanisms will open AI models' closed-box nature, offering model-agnostic and model-specific techniques [42]. Such methods could be divided into ex-ante and post-hoc techniques. While the first one incorporates interpretability directly into the AI model during its design and development stage, the latter provides insights into the AI model's process after it has been trained [164], [165]. Ex-ante techniques are proactive and are mainly built into the AI models from the start. Examples include AI models that are intrinsically interpretable, for instance, decision tree models or linear regression models. On the other side, post-hoc methods are applied after the AI models have been trained, so they could make AI predictions more interpretable. These might include visualization methods and feature importance scores that help end-users understand the

AI models' decisions. One option would be to provide the positive or negative predictive value of a finding.

3) ETHICAL FRAMEWORKS

Ethical frameworks offer comprehensive guidelines that build fairness and transparency throughout the design, development, and implementation of AI systems. These frameworks facilitate identifying ethical issues, including bias, discrimination, safety, and privacy concerns, enabling healthcare providers to proactively address such issues [166]. The ethical frameworks for responsible AI in healthcare take a multidisciplinary approach, integrating fields such as medical ethics, bioengineering, human well-being, regulatory compliance, and psychology into the AI development process. These frameworks highlight the importance of diverse expertise to not only address but also cope with complex ethical challenges. The responsible design process incorporates ethical impact analysis at each phase, including research, ideation, prototyping, and post-launch evaluation, focusing on psychological well-being and broader ethical issues such as social justice. For example, the ''Spheres of Technology Experience'' framework organizes the ethical impact of technology across six levels of adoption, interface, task, behavior, life, and society, allowing developers to evaluate and address potential ethical issues at every stage of user interaction and societal influence, making ethical analysis systematic, completely thorough, and actionable [167].

Ethical frameworks in healthcare emphasize the principles of measured action and caring in the in-between to support decision-making that recognizes the interconnectedness of stakeholders and the complexities of healthcare settings. ''Measured action'' involves taking small, adaptable steps in uncertain situations, while ''caring in the in-between'' focuses on making relationships among stakeholders and incorporating their concerns. Together, these principles help ensure the responsible and ethical integration of AI into healthcare, maximizing healthcare benefits while minimizing harm and preventing unequal treatment [95], [164], [166], [168], [169]. Moreover, ethical frameworks promote the protection of privacy by prioritizing patient consent for the use of their protected health information (PHI) and data minimization, as required by several regulations such as GDPR and HIPAA [95].

4) STAKEHOLDER ENGAGEMENT

Engaging stakeholders is fundamental to the successful establishment of REF-AI. Effective stakeholder engagement brings together AI scientists, healthcare providers, policymakers, patients, clinicians, health informatics professionals, IT specialists, and experts such as biomedical ethicists. This collaboration makes AI systems align well with ethical and societal standards, building a safe and sustainable AI ecosystem that benefits all stakeholders, end-users, and society as a whole [2], [99], [161], [163]. Ongoing collaboration among diverse stakeholders, including members of underrepresented and marginalized groups, can help reduce biases and associated risks. This approach leads to safer and trustworthy AI solutions that address the healthcare community's and society's needs [99], [163].

5) REGULAR AUDITS

Regular audits are key components of responsible AI in medical imaging, playing an important role in maintaining ethical standards, detecting and mitigating biases, and evaluating AI system performance against established benchmarks. These audits promote accountability and drive continuous improvement by systematically verifying and validating AI systems for adherence to ethical guidelines, legal frameworks, regulatory standards (e.g., the EU AI Act), and the specific requirements of medical imaging applications [164], [170].

Audits must be integrated throughout the entire AI life cycle, ranging from analysis of requirements and design through development, deployment, and routine utilization. A comprehensive auditing strategy should make the most use of internal and external evaluations conducted at all pre-deployment, post-deployment, and post-incident phases while maintaining audit independence to build an unbiased assessment [99], [170], [171], [172], [173]. Targeted data audits are essential for systematically evaluating the quantity, diversity, quality, availability, and integrity of data within AI systems. These audits help to first identify and then mitigate risks, such as data silos, biases, and integration challenges. By addressing these issues, data audits enhance the fairness, safety, transparency, and accountability of AI systems in medical imaging, building trust and equity in healthcare outcomes [2], [79], [173], [174].

6) ROBUST GOVERNANCE

Robust governance in responsible AI provides a well-organized, structured, and multi-tiered oversight framework that integrates risk management, regulatory compliance, ethical principles, and data governance. By operating across industry, organizational, and team levels, this governance structure helps in implementing AI systems that align well with societal values while advancing accountability, fairness, and transparency throughout their lifecycle [79], [158], [175], [176], [177]. This includes formal regulations, such as legislative acts and binding guidelines, with voluntary ethical principles to guide the responsible development, deployment, and maintenance of AI systems in healthcare. This dual approach not only enables organizations to achieve their long-term objectives in AI utilization but also safeguards stakeholder interests in AI-powered systems [111], [158], [175], [176], [178].

Effective AI governance, however, integrates several factors, such as risk management, regulatory compliance, and ethical considerations, into AI-driven decision-making processes to confirm the alignment of the AI methods with societal values. Furthermore, governance structures also

emphasize proactive oversight to keep AI-enabled systems accountable and adaptable to contemporary standards and societal expectations [177], [179], [180].

7) DATA PRIVACY

Data privacy is of paramount importance in the development of responsible AI systems. In responsible AI, data privacy includes establishing data governance structures that enforce data lineage, accountability, and adherence to privacy regulations, including HIPAA [181] and GDPR [182]. These frameworks incorporate metadata management through data catalogs and curation, enhancing transparency, integrity, and traceability. Such measures provide compliance with laws and help identify biases, validate data accuracy, and minimize risks associated with data integration, ultimately promoting fairness, accountability, and transparency across AI systems within the healthcare community [79], [183].

Advanced privacy-preserving techniques strengthen AI-powered systems by balancing security, privacy, availability, and efficiency. Federated learning, for instance, enables encrypted model parameters to be shared instead of raw data, maintaining user privacy by keeping data localized [109], [184]. Complementary approaches, such as secure multi-party computation, differential privacy, and trusted execution environments, could provide additional layers of protection [185], [186], [187]. These techniques facilitate computation on encrypted data instead of real raw and original data, ensuring compliance with privacy regulations (e.g., GDPR) without compromising the accuracy or utility of AI models [184], [188]. Moreover, robust data privacy practices must include informed consent protocols customized to diverse populations. Developing transparent and comprehensive consent is required to have individuals clearly understand how their data will be used, thus empowering them to make informed decisions for their data and how the data will be shared [189], [190].

8) USER-CENTRIC DESIGN

User-centric design in responsible AI emphasizes creating AI-powered technologies that prioritize user needs and enhance their overall experience with AI-enabled systems. This approach covers AI fairness, transparency, and ethical use while promoting accessibility and inclusivity for all individuals. By incorporating diverse societal values and adhering to universal design principles, user-centric design facilitates the development of AI systems that align with ethical standards and societal expectations [191], [192], [193]. In healthcare, a key aspect of user-centric design involves understanding and addressing clinicians' and patients' needs, preferences, and experiences through a focus on the interaction between all end-users and AI-enabled systems. This requires gathering user input, usability analysis, and evaluating systems from subjective perspectives. Different factors, such as AI transparency, fairness, privacy, and explainability, are central to this process. This helps to build and implement AI systems that not only meet technical requirements but also resonate with user expectations and values [191], [193], [194], [195]. Beyond software and system engineering disciplines, user-centric design incorporates insights from cognitive sciences, psychology, and the humanities to design and develop AI-powered systems that enhance human well-being and align with societal values. This multidisciplinary approach helps to incorporate fairness, transparency, and ethical use while addressing the broader impacts of AI technologies [192].

Moreover, accessibility is another critical pillar of user-centric design. It mainly aims to build AI-enabled systems that accommodate all users, regardless of their age, race, gender, abilities, or characteristics. Adhering to universal design principles may guarantee that AI systems are equitable, inclusive, and accessible to a broad range of end-users. By actively involving diverse cohorts, such as the aging population and individuals with disabilities in the development process, we will be able to implement AI strategies that uphold societal values, promote fairness, and support human dignity [193], [196], [197], [198], [199].

9) CONTINUOUS MONITORING

Continuous monitoring of responsible AI entails the ongoing validation and real-time observation of the functionalities and utilization of AI-enabled systems after deployment. This process maintains adherence to regulations, laws, and ethical standards while providing channels for a diverse range of stakeholders to provide feedback and report issues and comments [200], [201], [202], [203]. An important factor for the effectiveness of continuous monitoring is the engagement of a diverse user population, inclusive of different ages, races, genders, abilities, and ethnicities [162].

10) TRAINING AND EDUCATION

Training and education in responsible AI, delivered through courses, conferences, and multidisciplinary programs, provide continuous learning opportunities tailored to diverse audiences. These initiatives equip healthcare professionals, managers, and developers with essential knowledge of AI itself, AI ethics, regulation, clinical applications, and practical tools, thereby advancing trust, safety, and effective implementation of AI in healthcare settings [2], [8], [160], [204], [205]. In addition, training and education methods can improve responsible AI by equipping end-users, such as managers, developers, clinicians, and employees, with a strong understanding of AI ethics through structured programs, including but not limited to IEEE's initiative for ethical AI [204], [206], [207].

B. EXPLAINABLE AI

As discussed earlier, explainable AI in healthcare helps to open the closed-box nature of AI algorithms and makes it understandable for end-users, including clinicians, physicians, nurses, and patients. Generally speaking, explainable

AI methods could be categorized into (1) Attribution-based and (2) Non-attribution-based methods due to their distinct strategy in interpreting the decision-making processes of AI models (Figure 2). Attribution-based methods focus on quantifying the importance of individual input features, such as specific regions of interest, edges, and/or blobs in a medical image, and provide visual or non-visual representations to highlight these contributions. In contrast, non-attribution-based methods seek to uncover the broader mechanisms and reasoning behind an AI model's predictions or classification, focusing on the overall AI model behavior or high-level summaries that extend beyond individual features [40], [141], [208]. The current section introduces a taxonomy of explainable AI in medical imaging informatics.

1) ATTRIBUTION-BASED METHODS

The attribution-based methods in explainable AI are techniques used to enhance the interpretability of AI. The goal of these methods is to determine each feature's contribution to the target output/outcome. Attribution-based methods identify and highlight key features or regions within the input data, providing clearer insights into how the AI model arrives at their decisions [14], [141], [208], [209], [210]. These methods could then be classified into (1) visual attribution-based methods and (2) non-visual attribution-based methods. While the visual attribution-based methods aim to highlight the regions of the medical image that are contributing significantly to the AI model's prediction, the non-visual attribution-based methods mainly focus on identifying which features or internal model components (e.g., artificial neural network activations) are most responsible for an AI model's decision, without directly highlighting specific regions in the image.

a: VISUAL ATTRIBUTION-BASED METHODS

Class Activation Mapping: Class activation mapping (CAM) [211] is a widely used visualization technique that interprets deep CNNs by highlighting the image regions that are most important for making a specific class prediction. This makes it particularly effective for studies involving medical images [42], [212], [213], [214]. CAM employs global average pooling (GAP) after the final convolutional layers and before the fully connected layer, computing the class activation map by combining the weights of each convolutional filter with the activations at each spatial location. It highlights the regions most influential in the AI model's decision-making process, thereby visualizing which region of interest in the image is relevant for the AI model's prediction [42], [60]. CAM has been applied in various fields of medical imaging, including cancer identification and tumor classification, by visualizing salient regions in images of the bladder, brain, breast, skin, cardiovascular, chest, gastrointestinal, and thyroid scans [42], [60], [212], [215], [216], [217], [218], [219], [220], [221], [222], [223], [224], [225].

Gradient-Weighted Class Activation Mapping: Gradient-weighted class activation mapping (Grad-CAM) [34], [35] is a local explanation technique that addresses the limitations of CAM. It assigns importance scores to each artificial neuron by computing gradients flowing into the last convolutional layer, generating a coarse localization map, and highlighting key pixels for class prediction. Grad-CAM calculates artificial neuron importance scores by averaging gradients over the size of the activation map, then combines these weights with forward activation maps to produce a heatmap that highlights areas of the image most relevant for downstream tasks, such as segmentation or classification [41], [42], [52], [58], [61], [63], [226], [227], [228], [229]. Grad-CAM has diverse applications in medical imaging, such as visualizing regions significant for brain tumor detection, classifying polyps in whole slide images (WSI), identifying glaucoma in OCT scans, or pinpointing decision-critical areas in COVID-19 detection from chest radiographs [42], [60], [230], [231], [232], [233]. It also enhances visualization for polyp and tumor detection in endoscopic images and aids in gastric cancer classification by highlighting regions essential for diagnostic accuracy [58], [61], [62]. Grad-CAM's utilization extends to breast cancer imaging across different image modalities, such as ultrasound and mammography, where it effectively highlights lesion regions and assists clinicians in understanding any focusing areas [60], [63], [234], [235].

Moreover, Grad-CAM has proven highly effective in various medical imaging applications, enhancing AI model interpretability and diagnosis. It provides accurate heat maps for brain hemorrhage detection, aiding rapid diagnosis of hemorrhage locations [41]. In pneumonia detection, Grad-CAM improves the interpretability of X-ray images by focusing on relevant lung regions, even with background removal [227]. It is also valuable in lung CT imaging for localizing cancerous areas [43] and in brain tumor grading, where it highlights critical features like necrosis [53], [54]. Grad-CAM has been successfully adapted for tumor segmentation in MRI scans, including prostate cancer, offering performance like manual segmentation [228]. Additionally, it enhances radiograph and MRI interpretation in tasks like osteoarthritis diagnosis and tumor segmentation by emphasizing diagnostically relevant features [52].

Saliency Maps: Another mechanism under the category of visual attribution-based methods called saliency maps [236], [237]. Saliency maps are a gradient-based visualization technique that computes the impact of individual pixels on a neural network's classification. They do this by evaluating the gradients of the loss function concerning the input image, revealing which pixels most influence the final decision, and highlighting the relevance of different image areas for a given class [226]. This method highlights critical areas in medical images, helping researchers and clinicians understand where deep learning models focus and aiding in identifying and diagnosing potential issues or biases. Furthermore, saliency maps are used to explore how demographic factors, such

as race and sex, affect a deep learning model's predictions of brain regions associated with sex-linked neuropsychiatric conditions. By adding Gaussian noise and averaging the maps over multiple iterations, researchers improved clarity and identified key brain regions associated with different demographic subgroups [238].

Layer-Wise Relevance Propagation: Layer-wise relevance propagation (LRP) [239] is an interpretability method that explains the predictions of neural networks by redistributing the network's output back through the layers to assign relevance scores to individual input features, such as pixels in an image [42], [52], [226], [239], [240], [241]. Using local redistribution rules, LRP traces how much each artificial neuron contributes to the final output, producing a relevance map that highlights important regions of the input [42], [226], [242]. LRP has been widely used in medical imaging to identify key features for diagnosis. Examples include distinguishing schizophrenia patients using fMRI, creating heatmaps for tumor detection, diagnosing multiple sclerosis, and generating relevance maps for neonatal MRI and Alzheimer's disease detection [42], [241], [242]. Furthermore, LRP is also used to diagnose anterior disc displacement, osteoarthritis, and temporomandibular joint disorder from MRI images, generating heatmaps that serve as visualized rationales for diagnostic predictions [52], [240]. Those studies provide clinicians with interpretable insights and make artificial neural network models more transparent and interpretable for medical diagnosis.

Guided Back-Propagation: Guided back-propagation (GBP) [243] is a visualization and explainability technique that analyzes the gradient concerning the input image to highlight those features that are most influential to artificial neuron activation [53], [242]. It changes the backpropagation process by setting gradients to zero for units with zero or negative values after ReLU activation. It highlights features that increase activation and gives a clearer visualization than standard back-propagation [53], [226], [242]. This approach combines ReLU and deconvolution, introducing a guidance signal to prevent the backward flow of negative gradients. It makes it effective for visualizing artificial neural network activations in both the intermediate and final layers [226]. In medical imaging, GBP enables improved visualization and interpretation of diseases using neural networks. It aids in the automated quantification of enlarged perivascular spaces as markers of cerebral small vessel disease in brain MRI. For fMRI, it decodes task states of the human brain without feature engineering. GBP enhances the detection and visualization of bioresorbable scaffolds for coronary heart disease in intravascular optical coherence tomography (IVOCT). In colorectal imagery, it improves the semantic segmentation of polyps for cancer prevention with uncertainty estimation. For spinal MRI, it facilitates the grading and localization of pathologies like disc degeneration and stenosis with radiological evidence visualization [244], [245], [246], [247], [248], [249].

SmoothGrad: SmoothGrad [250] is an enhancement of gradient-based saliency maps that reduces noise by adding Gaussian noise to the input image and averaging the resulting sensitivity maps [42], [250], [251]. It improves the clarity of the saliency map by smoothing gradients, which reveals the effect of small changes in each pixel on the classification score. By applying a Gaussian kernel to average multiple perturbed images, SmoothGrad refines the saliency visualization and could be combined with other gradient-based methods for better results [42]. In medical imaging, SmoothGrad was applied to breast MRI data to enhance the clarity of feature visualizations generated by the deep CNNs, helping to distinguish relevant spatial and dynamic features from pre-processing artifacts during estrogen receptor status classification [251].

Occlusion Sensitivity: Occlusion sensitivity [236] is an agnostic method that generates saliency maps by systematically occluding parts of the input image and observing changes in the classification score to identify the importance of image regions in a model's decision [212], [236], [252], [253]. In medical image analysis, this technique generates heatmaps highlighting areas that significantly influence the model's decision, as demonstrated in lesion segmentation tasks. The method typically results in ''hotter'' explanation maps than GradCAM, suggesting more areas are considered highly relevant for segmenting masses [253], [254].

Integrated Gradients: Integrated gradients (IG) [255] is a gradient-based method used to interpret deep learning models by evaluating the contribution of each input feature to the model's prediction. It requires a baseline input, which could be a black/white or a random image, and calculates how the input image's features differ from the baseline to generate an explanatory heatmap [42], [240], [256]. IG satisfies the axioms of sensitivity and implementation invariance, making it a powerful tool for understanding feature importance and data skew [42], [255]. IG has been used in medical imaging to generate heatmaps for predicting diabetic retinopathy severity, providing pixel-level insights into feature contributions [257]. It has also visualized features in a deep CNN trained to classify estrogen receptor status from breast MRI [251] and helped create explainable heatmaps for diagnosing temporomandibular joint disc displacement using MRI images, supporting clinical decision-making [240].

Deep Learning Important Features: Deep learning important features (DeepLift) [258] is a back-propagation-based method that calculates contribution scores, such as LRP and IG, by comparing changes in artificial neuron activations between an input and a reference. By propagating these differences through the network, DeepLift quantifies how every feature contributes to the final prediction [42], [226]. It addresses issues with gradient-based methods, such as gradient zeroing and discontinuities, allowing for a more reliable interpretation of model decisions [42]. AI-powered Medical image analysis uses DeepLift to identify the salient features for Multiple Sclerosis classification, due to its

quantitative evaluation performance among other visually explainable AI methods [226], [259], [260].

Deconvolutional Networks: Deconvolutional networks (DeconvNets) [236] is an explainability technique used to visualize activations in CNNs by mapping pixel-level learning back to the input layer. This method constructs a deconvolution network by adding transposed convolution and unpooling layers to reverse the effects of convolution and max pooling layers in the original model. Deconvnets can effectively visualize activations for convolution and max pooling layers by generating visualizations that map learned features back to the input pixels [43].

b: NON-VISUAL ATTRIBUTION-BASED METHODS

Local Interpretable Model-Agnostic Explanation: Local interpretable model-agnostic explanation (LIME) [261] perturbs the original images, feeds them into the closed-box model, and examines the resulting outputs. It then assigns weights to these perturbed images based on their proximity to the original images [42], [55], [212], [229], [261]. It then fits a surrogate computational model, such as linear regression, to the weighted dataset of perturbed points, explaining the original image's prediction. LIME is applied to brain MRI data to reveal visual evidence supporting Alzheimer's disease classification using deep CNNs, and provides interpretable insights for medical professionals by pinpointing key brain regions that influence predictions [64], [262].

LIME generates heat maps and super-pixels to highlight key features influencing classification decisions for COVID-19 and pneumonia in CT and X-ray images, enhancing AI model explainability and aiding clinicians in understanding the decision-making process [263], [264]. It was also used to interpret deep learning predictions for retinoblastoma, identifying important regions in fundus images, improving transparency and detection accuracy [55]. Additionally, LIME provided plausible explanations for an endoscopic dataset including Lymphangiectasia and Pylorus using endoscopic images [229], and interpreted MRI patches for glioblastoma multiforme detection [265]. LIME has also improved the explainability of cancer detection models, such as those for colorectal cancer and osteosarcoma, by balancing accuracy with interpretability using histopathological images [266]. It further offers valuable visual explanations for CNN models in digital tomosynthesis images of breast lesion classification, highlighting influential regions [63].

SHapley Additive Explanations: SHapley Additive Explanations (SHAP) [267] is a game-theory strategy [268], [269] utilized to explain predictions by attributing the contribution of every input feature to the overall prediction of the AI model [55], [229], [270]. SHAP ensures local accuracy by aligning predictions with the expected average for simplified inputs, missingness by excluding features absent from the original input, and consistency by increasing the contribution of simplified inputs and the SHAP value if the model changes. This guarantees reliable feature attribution in AI model explanations [42]. In medical imaging, SHAP has been applied to analyze the output of a 3D regression CNN for estimating volumetric breast density from MRI images, identifying key features and highlighting inaccuracies related to structures like the pectoral muscle and heart, thus confirming the feasibility of estimating breast density without segmentation [270]. It was also used to enhance interpretability for retinoblastoma images by assigning importance scores to pixels, identifying key regions such as yellow-white masses and calcifications, and revealing the absence of critical features in normal images [55]. Additionally, SHAP was also employed to explain an AI model for the early detection of mutations in the Kirsten Rat Sarcoma viral oncogene homolog (KRAS) and epidermal growth factor receptor (EGFR) in lung cancer patients using low-dose computed tomography (LDCT) images [271].

Adversarially-Generated Counterfactuals: Adversarially-generated counterfactuals (ANCHOR) [272] is a method that provides stable explanations for model predictions by using a set of if-then rules to anchor a prediction to identify crucial features or segments in an image that consistently lead to a particular prediction. This method ensures that the variation among the remaining segments does not influence the prediction. In medical imaging, ANCHOR can generate a fairly intuitive explanation for a chest X-ray classification model for identifying COVID-19 patients by providing most of the left side of the lungs as the explanation for the prediction [42], [273].

2) NON-ATTRIBUTION-BASED METHODS

Non-attribution-based methods do not directly attribute importance to specific parts of the image; instead, they aim to reveal the underlying processes and rationale behind an AI model's predictions, offering explanations that go beyond pixel-level analysis [40], [208].

Counterfactual Explanations: Counterfactual explanations [274] are commonly employed in explainable AI to provide "what-if" insights, which explore how changes to the input data would affect the AI model's prediction. These explanations involve making minimal alterations to an image to interpret the AI model's decision for an individual instance [42]. In medical imaging, counterfactual explanations modify or exclude specific regions, such as pathological areas, to observe how predictions change. This approach helps identify key areas that influence the model's decision while preserving the interpretability of the model [275].

Testing With Concept Activation Vectors: Testing with concept activation vectors(TCAV) [276] is a non-visual global method in explainable AI that quantifies how high-level concepts, such as specific features like a tumor in a medical image, influence an AI model's predictions or decisions [42], [60], [277]. Several studies have applied TCAV to medical image analysis, including diabetic retinopathy (DR), cardiac MRI, breast cancer detection, and skin lesion

classification [42], [276], [278], [279], [280], [281]. More specifically, TCAV was used to identify significant diagnostic features like microaneurysms or aneurysms for DR and to assess clinically meaningful biomarkers, such as ventricular ejection in cardiac MRI [276], [278], [279]. Additionally, TCAV has been extended to regression problems and used in skin lesion classification to highlight important dermoscopic features [42], [277]. TCAV has also been used to evaluate how features, such as temperature gradients or vascular structures, affect classification decisions in infrared breast images [60]. TCAV with a discovering phase enhances AI interpretability in cardiac MRI by identifying key features linked to cardiac conditions, providing clinically meaningful explanations, and a quantitative measure of feature importance while reducing pre-processing time [278].

Prototype-Based: Prototype-based methods [282], [283], [284], [285] in AI involve representing data points using a set of prototype examples that capture the essence of different classes or decisions. Each prototype serves as a reference point for classifying new samples, mainly based on their similarity to these prototypes. Prototype-based methods involve using representative examples or prototypes to explain and interpret model predictions [286]. In medical imaging, studies have used prototype-based methods for various applications, including lesion classification, thyroid nodule diagnosis, cancer detection, and COVID-19 detection from chest X-rays [42]. In addition, ProtoTree, a Prototype-based method, is used in infrared breast image classification by employing a decision tree structure where each leaf represents a prototype for a class of images and classifies images based on their proximity to these prototypes [60]. Techniques, such as influence functions and variational autoencoders (VAEs), have been used to identify significant features, cluster image patches, and generate visual interpretations of prototypes, aiding in understanding and explaining model predictions and improving diagnostic accuracy [42].

C. FAIR AI

Fairness in AI could be classified into three different strategies, including pre-processing, in-processing, and post-processing strategies. These techniques are mainly based on when and how bias detection and mitigation techniques intervene in the AI model pipeline (Figure 2). Pre-processing methods involve organizing or adjusting the input data to mitigate bias before training an AI model. Common methods include data reweighting, which assigns different weights to samples to correct imbalances, and data anonymization, which removes sensitive attributes to prevent discrimination based on protected characteristics. Other methods, such as data resampling and adversarial debiasing, are also employed to ensure fairness in the model's performance [31], [59], [287], [288], [289], [290], [291], [292]. In contrast, in-processing strategies, such as adversarial training and fairness constraint methods, intervene during model training to mitigate bias and enforce fairness [31]. Post-processing strategies, on the other hand, adjust model outputs to calibrate predictions and ensure fairness across sub-groups [59].

1) PRE-PROCESSING

Group Rebalancing: Group rebalancing is a technique to address class or group bias in datasets to ensure that each group or class is represented more equally, improving model fairness and performance across diverse categories. Group rebalancing typically involves techniques, such as data resampling [288], data reweighting [293], stratified batching [294], and data augmentation [295]. Data resampling involves either oversampling minorities or undersampling majorities in order to ensure a balanced representation across the classes. In cardiac MR image analysis, the data resampling methods have been used to ensure that each batch of data includes an equal representation of all protected groups, thereby enhancing fairness in the AI model's training process [59], [296]. Data resampling and stratified batching have also been employed to address biases related to gender, race, and age for deep learning-based segmentation of the skeletal anatomy of the knee and hip joints in plain radiographs [28]. Data reweighting could cope with addressing label noise and imbalanced datasets in medical images by giving individual weights to training samples to reduce bias [292]. In medical images, data reweighting can improve model performance on noisy labeled data, such as skin lesion datasets, and can be applied to various medical image classification tasks without requiring pre-estimated noise distributions or clean data [292]. Data augmentation artificially boosts the diversity of training data (e.g., images) by generating variations of existing images, such as through rotations, flips, or noise addition [297]. Data augmentation has effectively reduced diagnostic accuracy disparities, such as those in diabetic retinopathy between different skin tones [31]. Moreover, the data augmentation method alters the retinal appearance and diabetic retinopathy status to address imbalances related to skin color in retinal images and improve fairness in medical image analysis [298]. It can also improve the performance and fairness of classifiers in histopathology, chest X-ray, and dermatology, especially for underrepresented groups and out-of-distribution cases [297].

Domain Generalization: Domain generalization (DG) [299] aims to maintain good performance across diverse and unseen sub-populations by addressing distribution shifts. DG methods include Group Distributionally Robust Optimization (GroupDRO), which reduces worst-case loss through stronger regularization, and Stochastic Weight Averaging Densely (SWAD), which improves model performance by finding robust flat minima through dense weight sampling. Another method, Sharpness-Aware Minimization (SAM), focuses on parameters in regions of consistently low loss to enhance generalization [296]. In medical imaging, DG is crucial because AI models trained on data from specific hospitals or populations might not generalize effectively to data from different sources or patient demographics.

DG techniques aim to improve model robustness and fairness by minimizing domain-specific biases and ensuring that the learned features represent the underlying medical conditions rather than the characteristics of a particular dataset.

2) IN-PROCESSING

Adversarial Training: Adversarial training [300] is a computational strategy to address biases by enhancing the primary model's performance on the target variable while preventing a secondary model from predicting sensitive attributes based on the primary model's features [31], [296]. Adversarial training has been shown to effectively reduce biases in skin lesion classification [31]. Adversarial training was also employed to predict HIV diagnosis from MRI scans, identify sex differences in adolescents using data from the National Consortium on Alcohol and Neurodevelopment in Adolescence (NCANDA), and identify bone age from plain radiographs [301]. This method can achieve accurate predictions while reducing biases related to confounder [301].

Fairness Constraint: Fairness constraint [302] is a mechanism incorporated into machine learning models to ensure that predictions are not influenced by sensitive attributes in the data, such as gender or race [303]. In medical imaging, fairness constraints can be applied to prevent disparities in diagnostic accuracy based on protected attributes, such as age, gender, race, and ethnicity to ensure equitable treatment across different patient demographic groups. More specifically, the fairness constraint has been utilized in artificial neural network architectures for dermatology medical image analysis, which involves the introduction of an unfairness score, which is the difference in precision between light and dark skin datasets, thus, it could help to minimize this score through targeted fairness constraints [304]. The fairness constraint is applied to a multiexit convolutional neural network (ME-CNN) to achieve fairness in the diagnosis of dermatological disease without using sensitive attributes. This addresses concerns about privacy and availability while improving discrimination based on low-level features and optimizing the accuracy and fairness balance for each test instance [305].

Fair Meta-Learning: Fair meta-learning is a method that trains a model to optimize both accuracy and fairness by introducing a meta-fair classifier into an AI-powered model [59]. For example, in tasks like cardiac MR image segmentation, this method has been used to handle image segmentation and classification of protected attributes in a multi-task learning framework. This approach aims to ensure that no single group disproportionately affects the learning of the model, thus promoting fairness between different groups of patients in medical imaging [59].

Subgroup-Tailored: Subgroup-tailored modeling [28], [306], [307] enhances model performance for minority groups by training individual AI models specific to each group, enabling evaluation of group-specific AI models' effectiveness in mitigating bias and improving overall fairness. In medical imaging, the subgroup-tailored modeling method, applied to automated image segmentation of knee and hip anatomy using plain radiographs, improved fairness by reducing racial biases [28]. This tailored strategy ensures that the AI model(s) perform well in diverse populations, addressing potential disparities, and improving the general fairness in medical image analysis.

3) POST-PROCESSING

Equalized Odds Post-Processing: Equalized odds post-processing (EOP) [308] corrects the output of an existing AI algorithm to satisfy equalized odds, a fairness measure requiring privileged and unprivileged groups to have identical false positive and false negative rates. It happens by solving a linear program to adjust output labels probabilistically [59], [71], [309]. EOP is applied to achieve group fairness in chest X-ray classifiers [65].

Calibrated Equalized Odds Post-Processing: Calibrated equalized odds post-processing [310] builds upon the equalized odds approach. However, it adjusts the output labels by optimizing classifier scores to probabilistically meet the equalized odds requirement. It mainly focuses on ensuring that the adjusted predictions are both well-calibrated and fair [59], [71], [309].

Reject Option Classification: Reject option classification [65] aims to reduce bias by adjusting predictions for cases where the AI model is uncertain. It works by giving preference to unprivileged groups and reducing the advantages for privileged groups within a specific confidence range around the decision boundary. This method uses probabilistic classifiers or ensemble classifier disagreements to make decisions in high-uncertainty situations without changing the original data or model. It also allows for flexible control over fairness, and it can handle multiple sensitive attributes at the same time [59], [71], [309].

IV. REF-AI EVALUATION METRICS

This section provides a brief overview of the evaluation metrics available for REF-AI.

A. RESPONSIBLE AI

Evaluating responsible AI requires qualitative and quantitative approaches, focusing on societal impacts, stakeholder engagement, robust governance, and adherence to ethical principles. This requires conducting iterative assessments within the AI lifecycle, ranging from requirement analysis to development and deployment. These assessments should integrate qualitative feedback from diverse groups, including clinicians, patients, policymakers, and marginalized communities, with quantitative analysis. In addition, methods such as surveys, participatory design evaluations, and ethical impact assessments are employed to ensure comprehensive and inclusive evaluation [2], [99], [311], [312], [313], [314], [315].

Unlike explainable and fair AI, which benefits from well-established metrics and measurement approaches,

TABLE 1. An overview of the key methodologies, strategies, and techniques categorized under the REF-AI framework.

Category	Method(s)	Reference(s)
Responsible AI	Bias Detection and Mitigation	(Norori et al., 2021) [159], (Sargent et al., 2021) [160], (Schwartz et al., 2022) [162], (Werder et al., 2022) [79], (Mensah et al., 2023) [161], (Cary et al., 2024) [163], (Siddiqui et al., 2024) [28], (Nasir et al., 2024) [95]
	Transparency Mechanisms	(Szabo et al., 2022) [165], (Diaz et al., 2023) [164], (Hossain et al., 2023) [42]
	Ethical Frameworks	(Peters et al., 2020) [167], (Li et al., 2022) [166], (Diaz et al., 2023) [164], (Valles et al., 2023) [168], (Bekbolatova et al., 2024) [169], (Nasir et al., 2024) [95]
	Stakeholder Engagement	(Siala et al., 2022) [2], (Gkontra et al., 2023) [99], (Mensah et al., 2023) [161], (Cary et al., 2024) [163]
	Regular Audits	(Werder et al., 2022) [79], , (Siala et al., 2022) [2], (Diaz et al., 2023) [164], (Gkontra et al., 2023) [99], (Xia et al., 2024) [170], (Esmaeilzadeh et al., 2024) [174], (Li et al., 2024) [171], (Murikah et al., 2024) [172], (Riemer et al., 2024) [173]
	Robust Governance	(Macrae et al., 2019) [158], (Ho et al., 2019) [178], (Guan et al., 2019) [180], (Reddy et al., 2020) [111], (Khanna et al., 2021) [179], (Werder et al., 2022) [79], (Camilleri et al., 2024) [175], (Ligot et al., 2024) [177], (Lu et al., 2024) [176]
	Data Privacy	(Yao et al., 1982) [185], (Garfinkel et al., 2003) [187], (Dwork et al., 2008) [186], (Yang et al., 2021) [184], (Ng et al., 2021) [109], (Kotsenas et al., 2021) [190], (Werder et al., 2022) [79], (morley et al., 2022) [183], (Rudd et al., 2023) [188], (Hurley et al., 2024) [189]
	User-Centric Design	(Dignum et al., 2019) [192], (Nevanpera et al., 2021) [193], (Oberste et al., 2022) [194], (H et al., 2023) [195], (Torkamaan et al., 2024) [191],
	Continuous Monitoring	(Nurhaliza et al., 2020) [203], (Sanderson et al., 2022) [201], (Lu et al., 2022) [202], (Schwartz et al., 2022) [162], (Radanliev et al., 2024) [200]
	Training and Education	(Chatila et al., 2017) [207], (Economou et al., 2019) [206], (Sargent et al., 2021) [160], (Walsh et al., 2023) [8], (Wang et al., 2020) [204], (Siala et al., 2022) [2], (Dominguez et al., 2023) [205]
Explainable AI	Class Activation Mapping (CAM)	(Zhou et al., 2016) [211], (Rajpurkar et al., 2018) [225], (Ausawalaithong et al., 2018) [222], (Shinde et al., 2019) [215], (Gao et al., 2019) [216], (Lee et al., 2019) [217], (Lee et al., 2019) [218], (Dunnmon et al., 2019) [223], (Luo et al., 2019) [220], (Lei et al., 2020) [219], (Wang et al., 2020) [221], (Li et al., 2020) [224], (Jung et al., 2021) [214], (Miro et al., 2022) [212], (Hossain et al., 2023) [42], (Raghavan et al., 2024) [60], (Wang et al., 2024) [213]
	Gradient-weighted Class Activation Mapping (Grad-CAM)	(Korbar et al., 2017) [230], (Selvaraju et al., 2017) [34], (Pereira et al., 2018) [53], (Thakoor et al., 2019) [232], (Lin et al., 2020) [233], (Vuppala et al., 2020) [43], (Windisch et al., 2020) [231], (Kim et al., 2021) [41], (Saleem et al., 2021) [58], (Karim et al., 2021) [52], (Zhang et al., 2021) [35], (Esmaeili et al., 2021) [54], (Javali et al., 2022) [62], (Yang et al., 2022) [227], (Hussain et al., 2022) [63] , (Gunashekar et al., 2022) [228], (Mukhtorov et al., 2023) [61],(Varam et al., 2023) [229], (Hossain et al., 2023) [42], (Raghavan et al., 2024) [60],(Lai et al., 2024) [226], (Kajala et al., 2024) [234], (Talaat et al., 2024) [235]
	Saliency Maps	(Simonyan et al., 2013) [237], (Zeiler et al., 2014) [236], (Stanley et al., 2022) [238], (Lai et al., 2024) [226]
	Layer-wise Relevance Propagation (LRP)	(Bach et al., 2015) [239], (Bohle et al., 2019) [242], (Yoon et al., 2023) [240], (Shin et al., 2023) [241], (Karim et al., 2021) [52], (Hossain et al., 2023) [42], (Lai et al., 2024) [226]
	Guided Back-Propagation (GBP)	(Springenberg et al., 2014) [243], (Jamaludin et al., 2017) [249], (Pereira et al., 2018) [53], (Dubost et al., 2019) [245], (Bohle et al., 2019) [242], (Eitel et al., 2019) [244], (Gessert et al., 2019) [247], (Wickstrom et al., 2020) [248], (Wang et al., 2020) [246], (Lai et al., 2024) [226]
	SmoothGrad	(Smilkov et al., 2017) [250], (Papanastasopoulos et al., 2020) [251], (Hossain et al., 2023) [42]
	Occlusion Sensitivity	(Zeiler et al., 2014) [236], (Fong et al., 2017) [252] (Miro et al., 2022) [212], (Farrag et al., 2023) [253], (Chen et al., 2024) [254]
	Integrated Gradients (IG)	(Sundararajan et al., 2017) [255], (Ancona et al., 2017) [256], (Sayres et al., 2019) [257], (Papanastasopoulos et al., 2020) [251], (Yoon et al., 2023) [240], (Hossain et al., 2023) [42]
	Deep Learning Important Features (DeepLift)	(Shrikumar et al., 2017) [258], (Hossain et al., 2023) [42], (Jin et al., 2023) [260], (Patil et al., 2024) [259], (Lai et al., 2024) [226]
	Deconvolutional Networks (DeconvNets)	(Zeiler et al., 2014) [236], (Vuppala et al., 2020) [43]
	Local Interpretable Model-Agnostic Explanation (LIME)	(Ribeiro et al., 2016) [261], (Rucco et al., 2020) [265],(Ahsan et al., 2021) [264], (Kamal et al., 2021) [262], (Hussain et al., 2022) [63], (Miro et al., 2022) [212], (Hossain et al., 2023) [42], (Duamwan et al., 2023) [64], (Ghnemat et al., 2023) [263], (Varam et al., 2023) [229], (Aldughayfiq et al., 2023) [55], (Alkhalaf et al., 2023) [266]
	SHapley Additive Explanations (SHAP)	(Von et al., 1947) [269], (Fudenberg et al., 1991) [268], (Lundberg et al., 2017) [267], (Van et al., 2020) [270], (Le et al., 2021) [271], (Hossain et al., 2023) [42], (Aldughayfiq et al., 2023) [55], (Varam et al., 2023) [229]
	Adversarially-Generated Counterfactuals (ANCHOR)	(Ribeiro et al., 2018) [272] (Abeyagunasekera et al., 2022) [273], (Hossain et al., 2023) [42]
	Counterfactual Explanations	(Wachter et al., 2017) [274], (Lenis et al., 2020) [275], (Hossain et al., 2023) [42]
	Testing with Concept Activation Vectors(TCAV)	(Kim et al., 2018) [276], (Clough et al., 2019) [279], (lucieri et al., 2020) [281], (Tjoa et al., 2020) [277], (Gamble et al., 2021) [280], (Janik et al., 2021) [278], (Hossain et al., 2023) [42], (Raghavan et al., 2024) [60]
	Prototype-based	(Li et al., 2018) [282], (Chen et al., 2019) [283] (Donnelly et al., 2022) [284], (Hossain et al., 2023) [42], (Bodria et al., 2023) [286] (Rath et al., 2024) [285], (Raghavan et al., 2024) [60]
Fair AI	Group Rebalancing	(Chawla et al., 2002) [288], (Kamiran et al., 2012) [293], (Xue et al., 2019) [292], (Shorten et al., 2019) [295],(Burlina et al., 2021) [298], (Puyol et al., 2021) [59], (Zong et al., 2022) [296], (Ricci et al., 2022) [31], (Afzal et al., 2023) [294], (Siddiqui et al., 2024) [28], (Ktena et al., 2024) [297]
	Domain Generalization	(Blanchard et al., 2011) [299], (Zong et al., 2022) [296]
	Adversarial Training	(Goodfellow et al., 2020) [300], (Zhao et al., 2020) [301], (Ricci et al., 2022) [31], (Zong et al., 2022) [296]
	Fairness Constraint	(Dwork et al., 2012) [302], (Zafar et al., 2019) [303], (Yang et al., 2023) [304], (Chiu et al., 2023) [305]
	Fair Meta-Learning	(Puyol et al., 2021) [59]
	Subgroup-Tailored	(Kearns et al., 2018) [306], (Feuerriegel et al., 2020) [307], (Puyol et al., 2021) [59], (Siddiqui et al., 2024) [28]
	Equalized Odds Post-processing	(Hardt et al., 2016) [308], (Lohia et al., 2019) [309], (Puyol et al., 2021) [59], (Zhang et al., 2022) [65], (Ferrara et al., 2023) [71]
	Calibrated Equalized Odds Post-processing	(Pleiss et al., 2017) [310], (Lohia et al., 2019) [309], (Puyol et al., 2021) [59], (Ferrara et al., 2023) [71]
	Reject Option Classification	(Lohia et al., 2019) [309], (Puyol et al., 2021) [59], (Zhang et al., 2022) [65], (Ferrara et al., 2023) [71]

responsible AI lacks widely accepted standardized metrics. Instead, it relies on qualitative evaluations and emerging frameworks that address diverse aspects such as accountability, transparency, robust governance, data privacy, and ethical compliance. The challenge lies in developing cohesive standards to measure the subjective dimensions of responsibility in AI [2], [8], [164], [176], [316], [317], [318], [319].

B. EXPLAINABLE AI

Subjective and objective metrics provide different approaches for evaluating AI explanations, each with its own methods and goals. Together, these metrics ensure human interpretability and AI model transparency (Figure 3). Here, we briefly discuss these metrics.

1) SUBJECTIVE METRICS

Subjective metrics include human evaluation to assess the quality and accuracy of AI explanations. These metrics rely on human judgment to assess how well an AI explanation is understood and how it influences trust among AI and end-users. This process often requires expert insights to determine if explanations align with human understanding and expectations. These metrics are often gathered through user studies or surveys, with evaluators assessing the relevance and correctness of AI explanations to provide more precise insights into the model's behavior and its decision-making processes [42], [320], [321].

2) OBJECTIVE METRICS

In contrast, objective metrics focus on measurable criteria (e.g., fidelity and consistency). These objective metrics in AI explainability could be divided into model-based and specific metrics.

Model-Based Metrics: These metrics focus on directly interacting with an AI model's predictions or processes, providing clear and measurable evaluations. Attribution-based metrics could check how well an AI explanation matches the AI model's importance assigned to different features. For example, they might measure the accuracy of saliency maps, the relevance of highlighted areas, or correlation and/or IoU among areas automatically detected by AI and those manually provided by the domain experts. Perturbation-based metrics look at how changes to the input or features affect the model's predictions, helping to understand the stability and reliability of the explanations. Additionally, model performance metrics assess how well the explanations reflect the model's behavior, often by tracking changes in accuracy or consistency when the inputs are altered [322], [323], [324], [325], [326], [327].

- *Attribution-Based Metrics:* To comprehensively evaluate attribution-based metrics, various methods could assess how accurately an AI explanation highlights relevant features so that key regions influencing model predictions are correctly identified and aligned with ground truth areas. These metrics could be categorized as follows.

  Pointing Game Metrics: Pointing game metrics evaluate the precision of the saliency maps in identifying the relevant regions [42], [212], [328]. Pointing game metric measures how well the maximum point from the saliency map aligns with the bounding box of a specific class. A hit is counted if the point falls within the bounding box; otherwise, it's a miss, and the accuracy is obtained from the number of misses and hits for each object [212].

  Attribution Localisation Metrics: Attribution Localisation metrics [329] are calculated as the ratio of the sum of positive attributions within the bounding box to the sum of overall attributions [330], [331].

  Top-k Intersection: Top-k intersection [332] evaluates the consistency of feature importance between different images. The top-k intersection measures the overlap of the top-k most relevant features between the original and perturbed image [42], [332].

  Concept Influence Score: This score assesses the relevance of high-level visual concepts in AI model predictions. The concept influence score measures the pixel-wise intersection between an explanation and a segmentation map, focusing on semantic concepts influencing predictions [42], [333].

  Relevance Mass Accuracy: Relevance mass accuracy [334] evaluates how well the ground truth mask captures the relevant features. It is calculated as the ratio of the sum of positive relevance values within the bounding box to the sum of all positive attributions in the image [42], [334], [335], [336].

  Relevance Rank Accuracy: Relevance rank accuracy [334] assesses the concentration of essential features within the relevant areas. It determines how high-intensity relevance is included within the ground truth mask [42], [334], [336], [337].

  Faithfulness Correlation: Faithfulness correlation [338] evaluates the faithfulness of the explanation to the model's predictions and measures the correlation between the explanation and the actual model behavior [42], [334], [335], [336], [337], [339].

- *Perturbation-Based Metrics:* Perturbation-based metrics assess how changes in features affect AI model predictions. By modifying specific input areas, pixels, or model parameters, these metrics test the stability and consistency of AI explanations by observing how the output and explanations change. These metrics could be classified as follows.

  Deletion and Insertion Metrics: Deletion and insertion metrics [340] examine the effect of removing or adding important pixels to observe changes in class probability. Deletion measures class probability decrease, when key pixels are removed, while insertion tracks probability increase with added pixels [42], [58], [340], [341], [342].

  Remove And Retrain: Remove and retrain (ROAR) [343] determines the effect of feature removal on model performance. ROAR involves perturbing the highest-scoring regions, retraining the AI model on these perturbed images, and checking for changes in accuracy [341], [342], [344].

  Remove and Debias: Remove and debias (ROAD) [345] reduces computational costs while evaluating feature importance. ROAD measures the impact of removing features without retraining, using mutual

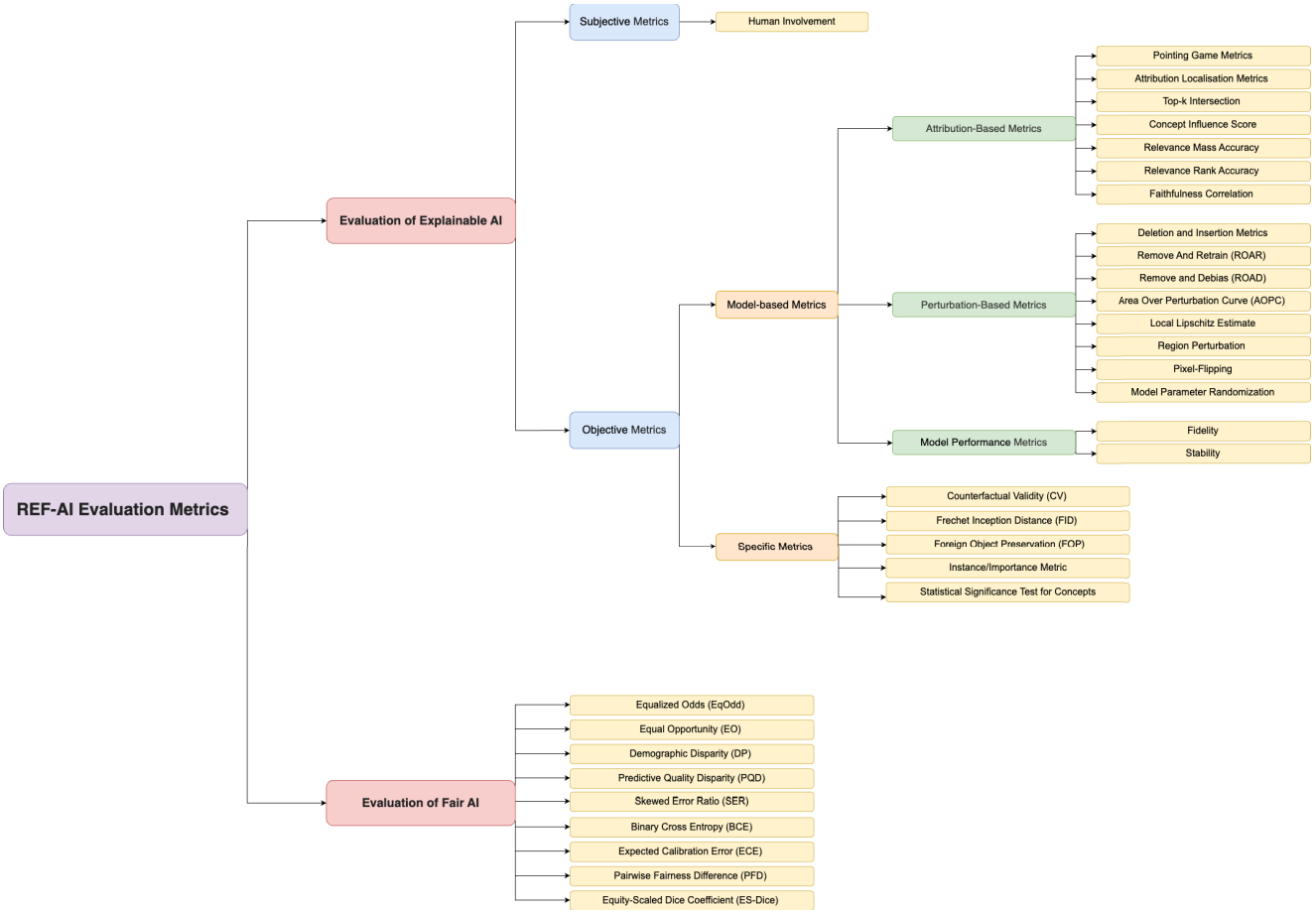


FIGURE 3. The taxonomy of the responsible, explainable, and fair artificial intelligence (REF-AI) evaluation metrics for medical image analysis.

information to assess the contribution of low-important pixels [42], [345].

Area Over Perturbation Curve: Area Over Perturbation Curve (AOPC) [340] quantifies the impact of perturbations on the saliency map's relevance. It measures the difference in certainty of object presence with and without perturbations and higher AOPC values indicate greater feature relevance [346], [347], [348].

Local Lipschitz Estimate: Local Lipschitz [349] Estimate assesses the stability of explanations across similar inputs and evaluates the consistency of explanations for similar instances using Lipschitz continuity [349], [350].

Region Perturbation: Region perturbation [340] measures the contribution of different regions to the model's output, and it perturbs specific regions in the input to evaluate their impact on the model's predictions [340], [351], [352], [353], [354].

Pixel-Flipping: Pixel-flipping [239] assesses the importance of specific pixels in the explanation, and it tests how flipping individual pixels affects the model's predictions [253], [355], [356].

Model Parameter Randomization: Model parameter randomization [357] assesses how explanations hold up under changes, and it evaluates the robustness of the explanation by randomizing model parameters and checking for consistency [42], [58], [228].

- *Model Performance Metrics:* Model performance metrics offer essential insights into the reliability and transparency of AI model explanations by evaluating how well these explanations align with the model's decision-making and maintain consistency under similar conditions. This includes two metrics as follows.

  Fidelity: Fidelity [358] measures how closely the methods resemble or mimic the decision-making process of the AI model, where high fidelity means that the AI explanation is a true representation of how the model makes its decisions [342], [358], [359], [360], [361].

  Stability: Stability [362] measures the consistency or coherence of explanations when handling similar instances, often calculated using the Lipschitz constant, which quantifies the sensitivity of AI explanations to input changes; a lower value indicates higher stability [42], [342].

Specific Metrics: These sorts of metrics target specialized evaluations, such as assessing counterfactual explanations

and concept learning models for their interpretability and adherence to human reasoning patterns. They could be classified as follows.

Counterfactual Validity: Counterfactual validity (CV) [363] ensures that the counterfactual explanation reflects a valid change in the prediction of the model and measures whether a counterfactual explanation corresponds to a change in the prediction of the classifier. If the classifier predicts an image as normal, the counterfactual should be classified as abnormal [364], [365].

Frechet Inception Distance: Frechet inception distance (FID) [366] evaluates how visually similar or different the counterfactual explanation is compared to the original input. FID quantifies the visual quality of counterfactual explanations by calculating the feature distance between the original input image and the counterfactual image [367], [368], [369], [370].

Foreign Object Preservation: Foreign object preservation (FOP) [366] ensures that the counterfactual explanation maintains the relevant details of the original input. FOP checks whether the counterfactual explanation retains individual patient information [66], [371].

Instance/Importance Metric: This will be divided into IM1 and IM2 metrics [372]. These metrics evaluate the consistency and quality of counterfactual explanations by analyzing reconstruction errors and similarities. While IM1 calculates the reconstruction error ratio between counterfactual instances, IM2 compares similarities among reconstructed counterfactual instances [364], [373], [374], [375], [376].

Statistical Significance Test for Concepts: A statistical significance test for concepts helps avoid false results by ensuring that the concepts used are important and relevant to the class prediction. It checks the stability and importance of concept activation vectors in concept-based models. A statistical test, like a two-sided t-test, is used to see if the concept activation vectors are strongly related to class predictions [276], [377], [378], [379], [380], [381], [382].

C. FAIR AI

Selecting appropriate fairness metrics is essential for evaluating equity in AI-enabled medical imaging models, as these metrics help identify and quantify biases, preventing AI systems from disproportionately disadvantaging specific groups. Researchers have developed a variety of fairness metrics to gain valuable insights into AI model performance, including demographic parity, equal opportunity, and predictive quality disparity [383], as illustrated in Figure 3). This section provides an overview of AI fairness metrics, with particular emphasis on their application in AI-powered medical imaging.

Equalized Odds: Equalized odds (EqOdd) [308] is a fairness metric that requires both the True Positive Rate (TPR) and the False Positive Rate (FPR) to be equal across different sensitive and/or protected groups (e.g., sex, age, race). This means that the model should make correct predictions (True Positives) and incorrect predictions (False Positives) at the same rate for each group, ensuring no group is disadvantaged in either aspect [296], [384], [385], [386].

Equal Opportunity: Equal opportunity (EO) [308] is a simpler version of Equalized odds. It focuses on equalizing the True Positive Rate (TPR) across different sensitive and/or protected groups. However, it does not require the rate of incorrect positive predictions (False Positive Rate or FPR) to be the same across groups [304], [387], [388], [389], [390].

Demographic Disparity: Demographic disparity (DP) [302] is a fairness metric that quantifies the percentage difference in positive outcomes across different demographic groups, measuring the diversity of positive outcomes for each sensitive and/or protected group [304], [391], [392], [393].

Predictive Quality Disparity: Predictive quality disparity (PQD) [383] measures the disparity in prediction quality among subgroups by calculating the ratio between the lowest and highest accuracy across those groups [304], [305], [383], [394].

Skewed Error Ratio: Skewed error ratio (SER) [395] evaluates fairness by measuring the skew or imbalance in error rates, such as false positives and false negatives. These are measured across specific groups by calculating the ratio of the highest to lowest error rate among those groups, with higher values indicating more significant bias and values closer to one reflecting lower bias [28], [59], [395], [396], [397].

Binary Cross Entropy: Binary cross entropy (BCE) is a loss function used to measure how well the predicted AI model's outcomes in a binary fashion align with the actual binary outcomes [65], [296], [398].

Expected Calibration Error: Expected calibration error (ECE) [399], [400] measures how well predicted probabilities align with true outcomes, indicating group sufficiency. While lower ECE confirms consistent threshold performance across different groups, higher ECE values may suggest a need for different optimal thresholds [65], [296], [310], [401].

Pairwise Fairness Difference: Pairwise fairness difference (PFD) [388] measures fairness by comparing the differences in AI model's outcomes between pairs of subgroups, where a large PFD indicates significant disparities and a lack of fairness in the AI model's predictions [402], [403], [404], [405].

Equity-Scaled Dice Coefficient: Equity-scaled dice coefficient (ES-Dice) adjusts the Dice coefficient to account for performance disparities across subgroups, providing a straightforward evaluation and more interpretable measure of fairness by comparing overall and subgroup-specific Dice coefficients [67], [406], [407].

V. CASE STUDY: HIP AND KNEE BONY ANATOMY SEGMENTATION USING REF-AI

AI-powered segmentation of hip and knee bony anatomy is essential for pre-surgical planning, prosthesis design, and the evaluation of musculoskeletal disorders. However,

TABLE 2. The summary of REF-AI evaluation metrics.

Category	Evaluation Metric	Reference(s)
Explainable AI	Human Involvement	(Doshi et al., 2017) [320], (Muddamsetty et al., 2021) [321], (Hossain et al., 2023) [42]
	Pointing Game Metrics	(Zhang et al., 2018) [328], (Miró-Nicolau et al., 2022) [212], (Hossain et al., 2023) [42]
	Attribution Localisation Metrics	(Kohlbrenner et al., 2020) [329], (Papanastasiou et al., 2023) [331], (Rio et al., 2024) [330]
	Top-k Intersection	(Ghorbani et al., 2019) [332], (Hossain et al., 2023) [42]
	Concept Influence Score	(Theiner et al., 2022) [333], (Hossain et al., 2023) [42]
	Relevance Mass Accuracy	(Arras et al., 2022) [334], (Srinivasan et al., 2022) [336], (Springenberg et al., 2023) [335], (Hossain et al., 2023) [42]
	Relevance Rank Accuracy	(Arras et al., 2022) [334], (Srinivasan et al., 2022) [336], (Wickstrom et al., 2023) [337], (Hossain et al., 2023) [42]
	Faithfulness Correlation	(Bhatt et al., 2020) [338], (Arras et al., 2022) [334], (Srinivasan et al., 2022) [336], (Komorowski et al., 2023) [339], (Wickstrom et al., 2023) [337], (Springenberg et al., 2023) [335], (Hossain et al., 2023) [42]
	Deletion and Insertion Metrics	(Samek et al., 2016) [340], (Saleem et al., 2021) [58], (Hu et al., 2022) [341], (Hossain et al., 2023) [42], (Nielsen et al., 2023) [342]
	Remove And Retrain (ROAR)	(Hooker et al., 2019) [343], (Hu et al., 2022) [341], (Nielsen et al., 2023) [342], (Sadafi et al., 2023) [344]
	Remove and Debias (ROAD)	(Rong et al., 2022) [345], (Hossain et al., 2023) [42]
	Area Over Perturbation Curve (AOPC)	(Samek et al., 2016) [340], (Kallipolitis et al., 2023) [347], (Gallo et al., 2023) [348], (Lamprou et al., 2024) [346]
	Local Lipschitz Estimate	(Alvarez et al., 2018) [349], (Doumard et al., 2023) [350]
	Region Perturbation	(Samek et al., 2016) [340], (Nam et al., 2020) [354], (Sun et al., 2023) [351], (Binder et al., 2023) [352], (Kadir et al., 2023) [353]
	Pixel-Flipping	(Bach et al., 2015) [239], (Pitroda et al., 2021) [356], (Farrag et al., 2023) [253], (Gnanavel et al., 2024) [355]
	Model Parameter Randomization	(Adebayo et al., 2018) [357], (Saleem et al., 2021) [58], (Gunashekar et al., 2022) [228], (Hossain et al., 2023) [42]
	Fidelity	(Guidotti et al., 2019) [358], (Velmurugan et al., 2021) [359] (Huang et al., 2021) [361], (Nielsen et al., 2023) [342], (Buliga et al., 2024) [360]
	Stability	(Alvarez et al., 2018) [362], (Hossain et al., 2023) [42], (Nielsen et al., 2023) [342]
	Counterfactual Validity (CV)	(Mothilal et al., 2020) [363], (Verma et al., 2020) [364], (Verma et al., 2024) [365]
	Frechet Inception Distance (FID)	(Ghandeharioun et al., 2021) [370], (Singla et al., 2023) [366], (Lamiable et al., 2023) [367], (koetzier et al., 2024) [368], (Xu et al., 2024) [369]
	Foreign Object Preservation (FOP)	(Singla et al., 2023) [366], (Patricio et al., 2022) [371], (Patricio et al., 2023) [66]
	Instance/Importance Metric	(Mahajan et al., 2019) [373], (Verma et al., 2020) [364], (Van et al., 2021) [372], (Khorram et al., 2022) [376], (Labaien et al., 2023) [375], (Taylor et al., 2024) [374],
	Statistical Significance Test for Concepts	(Kim et al., 2018) [276], (Ghorbani et al., 2019) [377], (Yeh et al., 2020) [378], (Foscarin et al., 2022) [380], (Ma et al., 2024) [379], (Kowal et al., 2024) [381], (Chanda et al., 2024) [382]
Fair AI	Equalized Odds (EqOdd)	(Hardt et al., 2016) [308], (Goel et al., 2018) [385], (Awasthi et al., 2020) [384], (Jung et al., 2021) [386] (Zong et al., 2022) [296]
	Equal Opportunity (EO)	(Hardt et al., 2016) [308], (Beutel et al., 2019) [389], (Narasimhan et al., 2020) [388], (Saha et al., 2020) [387], (Huang et al., 2022) [390], (Yang et al., 2023) [304]
	Demographic Disparity (DP)	(Dwork et al., 2012) [302], (Castelnovo et al., 2022) [391], (Yang et al., 2023) [304], (Ghosh et al., 2022) [393], (Cohausz et al., 2024) [392]
	Predictive Quality Disparity (PQD)	(Du et al., 2020) [383], (Du et al., 2022) [394], (Chiu et al., 2023) [305], (Yang et al., 2023) [304]
	Skewed Error Ratio (SER)	(Wang et al., 2020) [395], (Wang et al., 2020) [396], (Puyol-Antón et al., 2021) [59], (Siddiqui et al., 2024) [28], (Ohki et al., 2024) [397], (Atzori et al., 2024) [398]
	Binary Cross Entropy (BCE)	(Ruby et al., 2020) [399], (Zong et al., 2022) [296], (Zhang et al., 2022) [65]
	Expected Calibration Error (ECE)	(Guo et al., 2017) [400], (Pleiss et al., 2017) [310], (Nixon et al., 2019) [401], (Zong et al., 2022) [296], (Zhang et al., 2022) [65], (Brahmbhatt et al., 2023) [402]
	Pairwise Fairness Difference (PFD)	(Narasimhan et al., 2020) [388], (Hazirbas et al., 2021) [404], (Lin et al., 2023) [403], (Dodevska et al., 2023) [406], (Ferrara et al., 2024) [405]
	Equity-Scaled Dice Coefficient (ES-Dice)	(Tian et al., 2023) [407], (Masroor et al., 2024) [408], (Tian et al., 2025) [67]

challenges such as variations in patient anatomy, differences in imaging modalities, and demographic diversity raise important questions about the need for responsible and equitable AI solutions. As AI rapidly grows in healthcare, developing sophisticated AI systems that can address these challenges and improve equitable AI utilization for all patient groups becomes increasingly important. Furthermore, without adequate explainability, the adoption of AI in healthcare may be slowed, limiting its potential to improve outcomes across diverse populations. That being said, the explainability of AI models is important to build trust and transparency in clinical settings, enabling healthcare providers to understand and validate AI-driven insights with a high level of confidence [408].

This case study explores how REF-AI principles can address these challenges by integrating fairness, transparency, and accountability into AI-driven image segmentation techniques. By leveraging insights and methodologies from our recent work in the literature [28], we aim to demonstrate how the application of REF-AI principles improves technical performance and embeds ethical considerations into the clinical use of AI.

A. METHODOLOGY

The pipeline for REF-AI in bony anatomy segmentation includes the following:

1) DATA GOVERNANCE

Data was sourced from the Osteoarthritis Initiative (OAI) dataset [409], a diverse imaging repository that includes varying demographics, imaging machines, and clinical sites,

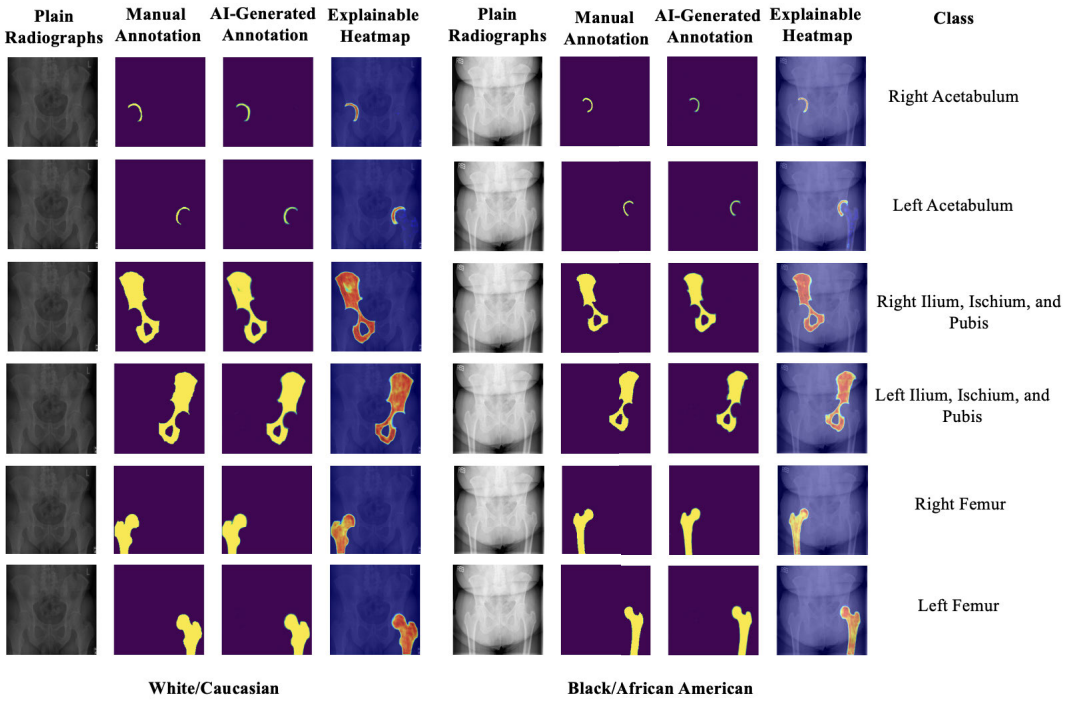


FIGURE 4. Generated Grad-CAM heatmaps for hip bony anatomy segmentation across the protected attribute race.

ensuring equitable representation of medical images. The OAI dataset is a longitudinal study that provides imaging such as plain knee radiographs, MRIs, and associated clinical outcomes. The OAI dataset's design includes standardized imaging protocols across multiple clinical sites, ensuring consistency and high-quality data collection. Furthermore, demographic diversity was carefully accounted for, with participants representing various ages, sexes, and ethnic backgrounds, allowing for equitable AI training and evaluation. OAI's data governance adhered to established privacy frameworks, including HIPAA, ensuring patient confidentiality. Comprehensive metadata documentation supported data traceability, enabling reproducibility and regulatory compliance. These qualitative and quantitative measures ensure that the dataset meets the ethical and technical requirements to develop fair and reliable segmentation models.

2) ALGORITHM DESIGN

To develop advanced AI-powered segmentation models, we utilized a U-Net architecture with a pretrained ResNet18 backbone, leveraging transfer learning to improve feature extraction. To address class imbalances, we applied data balancing techniques such as undersampling of overrepresented classes. Stratified sampling was used to reflect the demographic diversity of the population in the training and testing datasets, particularly in terms of sex and race. Furthermore, we trained specialized AI models for specific protected attributes, enabling us to evaluate group-specific [28], [410] performance and implement targeted bias mitigation strategies. These strategies were designed to make the AI-powered segmentation models equitable and reliable, providing accurate results across diverse patient populations.

3) AI EXPLAINABILITY USING GRADCAM

Grad-CAM [34], [35] was utilized to provide visual explanations of the segmentation process, enabling clinicians to understand the AI model. Grad-CAM highlighted regions of interest in the imaging data that most influenced the model's predictions, offering clear transparency into the decision-making process. Grad-CAM visualizations, shown in Figures 4 and 5 demonstrate the potential to align AI outputs with clinician expectations and build trust in the system.

4) FAIRNESS EVALUATION

The fairness of the various strategies was evaluated using the skewed error ratio (SER) [395]. This metric quantifies bias towards protected attributes by measuring the disparities in the model's prediction errors. SER is calculated by dividing the highest error rate to the lowest error rate among the protected or sensitive groups. A higher SER value indicates greater bias, while values closer to one reflect minimal bias and greater fairness in model performance.

B. RESULTS

The application of the REF-AI framework to hip and knee bony anatomy segmentation yielded promising results. By integrating diverse demographic data from the OAI

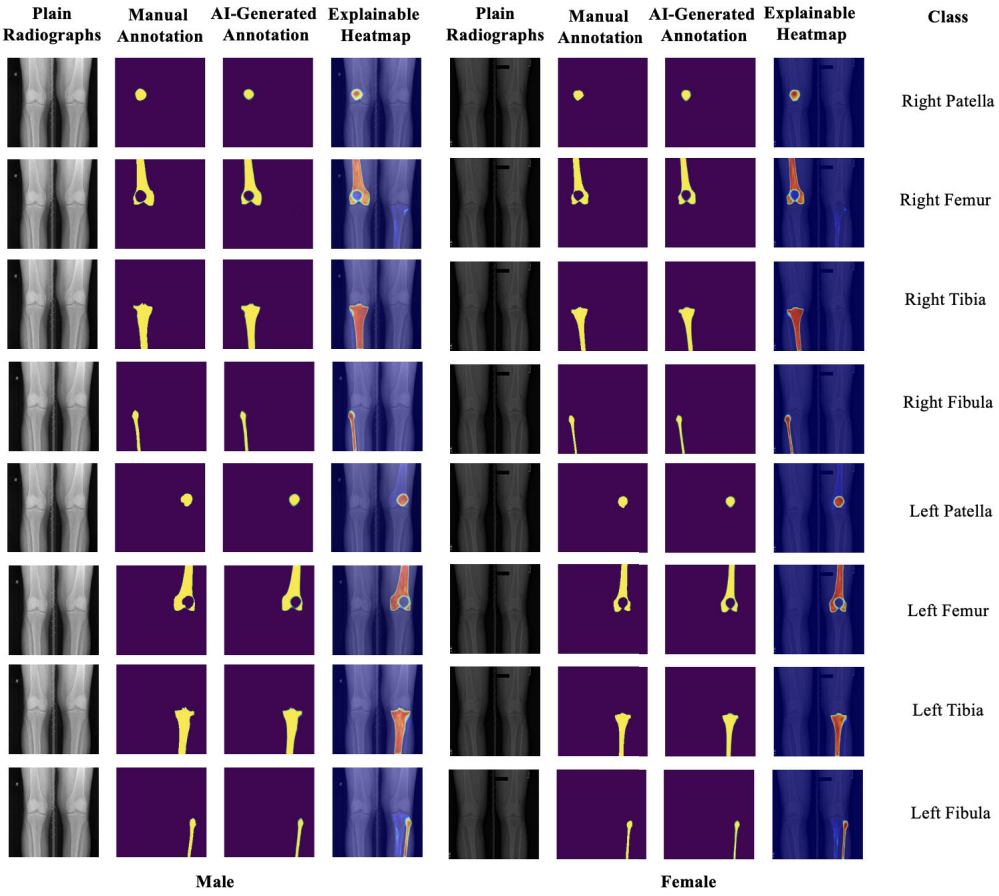


FIGURE 5. Generated Grad-CAM heatmaps for knee bony anatomy segmentation across the protected attribute sex.

dataset [409], the study achieved broad representation and equitable model performance. The advanced segmentation models demonstrated robust performance in identifying bony structures across varying patient demographics. Bias mitigation strategies minimized disparities across protected demographic groups, as indicated by the SER values in Tables 3 - 6. Furthermore, Grad-CAM visualizations, shown in Figures 4 and 5, successfully highlighted key anatomical landmarks. These visualizations aligned the AI model outputs with clinician expectations, fostering trust in the system.

Analysis of Tables 3 - 6 reveals key trade-offs between segmentation accuracy (measured by IoU) and fairness (evaluated using SER). Specifically, incorporating fairness into the segmentation algorithm often reduces performance for either one or both protected attribute groups. These trade-offs can be seen across all the results. For example, in Table 5, the Baseline model achieves higher IoU scores but demonstrates a higher disparity in fairness, with an SER of 1.112 across racial groups. On the other hand, while improving fairness (SER=1.050), the group-specific strategy [28] results in a reduction in IoU for the protected subgroups, demonstrating that achieving equitable performance often necessitates compromising overall segmentation precision.

Furthermore, the results demonstrate that no single bias mitigation strategy is universally optimal across all protected groups. For example, for hip segmentation, the Balanced model performs best for sex, whereas the Group-specific model is optimal for race. In contrast, for knee segmentation, the Stratified model yields the best performance for both sex and race.

C. CASE STUDY HIGHLIGHT

This case study demonstrates the significance of integrating REF-AI principles in medical imaging. The results emphasize the value of comprehensive data governance, with OAI's standardized imaging protocols and privacy frameworks ensuring high-quality, ethically compliant datasets. Targeted bias mitigation strategies significantly addressed disparities related to demographic diversity, including variations in sex and race attributes. These efforts enhanced fairness and contributed to more inclusive and representative AI model performance. Moreover, AI explainability tools such as Grad-CAM provided critical insights into the model's decision-making process. These scientific visualizations allow researchers and clinicians to better understand how the

TABLE 3. IoU and fairness scores for hip segmentation across the protected attribute sex.

Model	Male IoU	Female IoU	SER	SD
Baseline	0.870	0.868	1.014	0.000
Balanced	0.851	0.852	1.007	0.000
Stratified	0.870	0.862	1.054	0.000
Group-specific	0.853	0.847	1.044	0.002

TABLE 4. IoU and fairness scores for knee segmentation across the protected attribute sex.

Model	Male IoU	Female IoU	SER	SD
Baseline	0.924	0.925	1.021	0.001
Balanced	0.899	0.909	1.107	0.005
Stratified	0.921	0.920	1.015	0.001
Group-specific	0.916	0.920	1.050	0.002

TABLE 5. IoU and fairness scores for hip segmentation across the protected attribute race.

Model	White/Caucasian IoU	Black/AA[1] IoU	SER	SD
Baseline	0.875	0.861	1.112	0.007
Balanced	0.856	0.847	1.062	0.004
Stratified	0.864	0.849	1.102	0.007
Group-specific	0.853	0.846	1.050	0.003

TABLE 6. IoU and fairness scores for knee segmentation across the protected attribute race.

Model	White/Caucasian IoU	Black/AA[2] IoU	SER	SD
Baseline	0.924	0.925	1.015	0.000
Balanced	0.902	0.921	1.229	0.009
Stratified	0.916	0.916	1.000	0.000
Group-specific	0.919	0.915	1.055	0.002

AI model arrived at its predictions, enhancing transparency and alignment with clinical reasoning. This transparent approach builds trust among healthcare professionals and bridges the gap between AI-driven insights and real-world clinical applications.

Overall, this case study illustrates how the thoughtful application of REF-AI principles can enhance the technical robustness and ethical integrity of AI-powered medical imaging solutions, ultimately improving patient and clinical outcomes and equity in healthcare.

VI. DISCUSSION AND OUTLOOK

This paper explores the integration of responsibility, explainability, and fairness in AI, particularly within medical image analysis. A few years ago, the primary focus in AI development was mainly on improving AI model accuracy. However, as AI continues to play an increasing role in healthcare, it is clear that the conversation should be expanded. Today, accuracy for AI algorithms alone is no longer sufficient, meaning that the success of AI models in healthcare hinges on their performance combined with ethical considerations, especially in responsibility, explainability, and fairness. These factors are now essential for deploying AI technologies in a way that is effective and socially acceptable.

The AI fairness strategies, including pre-processing, in-processing, and post-processing, each offer distinct advantages, but their real-world application remains challenging. Pre-processing methods, such as group rebalancing and data augmentation, help mitigate biases in datasets before the AI model is trained, but they must be carefully calibrated to avoid overfitting and enhance the AI model's generalizability. In-processing strategies, such as adversarial training and fairness constraints, adjust the learning process to build equitable predictions. However, they often increase computational costs and, in some cases, may impact AI model accuracy. Post-processing techniques, such as equalized odds and reject option classification, allow for the calibration of model outputs after training to fine-tune fairness across various protected and/or sensitive subgroups. However, these methods alone cannot fully fix potential biases in the data and AI, thus, a more complete and systematic approach to fairness is needed. Fairness in AI will not be obtained in isolation. The challenges of mitigating biases related to sex, race, ethnicity, age, economic situation, and level of healthcare education persist in medical imaging, and efforts to enhance fairness must be balanced with the need for AI model performance. This balance is particularly challenging in healthcare, where the stakes are high and any compromise in AI model accuracy can have profound consequences. Furthermore, the need for fairness must be considered in the context of broader societal values, which can vary significantly across different cultures, states, and community settings. This underscores that addressing fairness in AI is not merely a technical task; it is, however, an ethical imperative that requires continuous collaboration and dialogue among diverse stakeholders.

Explainability also plays a significant role in the responsible use of AI systems. As AI becomes more popular in healthcare systems, it is essential that the AI models are transparent and their decision-making processes are easy to understand. This will help to build trust among all end-users. The evaluation of AI explainability must consider both subjective and objective metrics. Subjective metrics ensure that AI explanations are aligned with human understanding, while objective methods, such as attribution-based, offer a more quantitative understanding of how models make decisions. However, these metrics must be applied carefully to avoid oversimplifying or overcomplicating complex AI models.

Responsibility in AI extends beyond fairness and explainability; it requires ethical principles that guide the entire AI lifecycle. Developing standardized metrics for responsible AI remains a significant challenge. Current frameworks often focus on qualitative evaluations, such as stakeholder engagement and ethical compliance, but there is a need for more solid and universally applicable standards. One promising direction is the integration of fairness, accountability, and transparency frameworks, which could help establish

[1]African American.
[2]African American.

consistent guidelines for responsible AI across different healthcare domains. In medical imaging, where the impact of AI on human lives is profound, creating universally accepted standards that balance fairness, accountability, and transparency will be significant for guiding ethical decision-making.

Building truly responsible, explainable, and fair AI (REF-AI) in medical imaging informatics will require a collaborative and multidisciplinary framework that engages all parties and users. This includes technical experts, developers, AI programmers, patients, caregivers, healthcare providers, clinicians, surgeons, physicians, nurses, policy-makers, ethicists, and other relevant groups. Such collaboration is essential to implementing AI systems that reflect diverse needs, values, and ethical considerations.

While our study provides a solid review of the current advancements in REF-AI for medical imaging, it also carries some limitations. First, it focuses specifically on responsible, explainable, and fair AI (REF-AI) in medical imaging, which may limit its applicability to broader AI applications in healthcare. While the review discusses various REF-AI methodologies, it does not empirically validate or benchmark their effectiveness, making it difficult to assess their real-world performance. Furthermore, the field of AI governance and ethical AI deployment is rapidly evolving, and some of the legal or ethical considerations discussed may become outdated as new regulations emerge. Another limitation is that the review primarily addresses technical aspects of REF-AI but does not deeply explore its practical implementation challenges in clinical workflows.

In conclusion, the intersection of responsibility, explainability, and fairness is essential for developing AI systems that are ethical, transparent, and trustworthy, particularly in high-risk domains, such as medical image analysis. While the advancement of AI fairness offers promising solutions to detect and mitigate biases, their real-world application requires careful consideration of trade-offs between fairness and AI model performance. The role of explainability is similarly critical, helping to build AI systems that are transparent and aligned with human understanding. Despite significant progress, challenges remain in harmonizing these three pillars of AI responsibility, explainability, and fairness in complex domains in healthcare. Further research should prioritize refining these strategies, addressing the ethical dilemmas associated with their application, and establishing a standardized framework. This framework would be instrumental in developing comprehensive metrics for the design, development, evaluation, and implementation of REF-AI, ensuring that these systems are both effective and ethically sound.

ACKNOWLEDGMENT

The authors express their sincere gratitude to Cynthia Wycroft and Wendy Hoyt for their invaluable assistance in the preparation of this manuscript.

REFERENCES

[1] V. Dignum, "Responsibility and artificial intelligence," *Oxford Handbook Ethics AI*, vol. 4698, p. 215, Aug. 2020.
[2] H. Siala and Y. Wang, "SHIFTing artificial intelligence to be responsible in healthcare: A systematic review," *Social Sci. Med.*, vol. 296, Mar. 2022, Art. no. 114782.
[3] L. Alkire, A. Bilgihan, M. Bui, A. J. Buoye, S. Dogan, and S. Kim, "RAISE: Leveraging responsible AI for service excellence," *J. Service Manage.*, vol. 35, no. 4, pp. 490–511, Jul. 2024.
[4] J. Everson, J. Smith, K. Marchesini, and M. Tripathi, "A regulation to promote responsible AI in health care," *Health Affairs Forefront*, vol. 2024, Feb. 2024.
[5] C. Trocin, P. Mikalef, Z. Papamitsiou, and K. Conboy, "Responsible AI for digital health: A synthesis and a research agenda," *Inf. Syst. Frontiers*, vol. 25, no. 6, pp. 2139–2157, Dec. 2023.
[6] F. Bildirici, "Open-source AI: An approach to responsible artificial intelligence development," *REFLEKTİF Sosyal Bilimler Dergisi*, vol. 5, no. 1, pp. 73–81, 2024.
[7] M. T. Contaldo, G. Pasceri, G. Vignati, L. Bracchi, S. Triggiani, and G. Carrafiello, "AI in radiology: Navigating medical responsibility," *Diagnostics*, vol. 14, no. 14, p. 1506, Jul. 2024.
[8] G. Walsh, N. Stogiannos, R. Van De Venter, C. Rainey, W. Tam, S. McFadden, J. P. McNulty, N. Mekis, S. Lewis, T. O'Regan, A. Kumar, M. Huisman, S. Bisdas, E. Kotter, D. P. Dos Santos, C. Sá Dos Reis, P. Van Ooijen, A. P. Brady, and C. Malamateniou, "Responsible AI practice and AI education are central to AI implementation: A rapid review for all medical imaging professionals in Europe," *BJR Open*, vol. 5, no. 1, Aug. 2023, Art. no. 20230033.
[9] C. Mello-Thoms and C. A. B. Mello, "Clinical applications of artificial intelligence in radiology," *Brit. J. Radiol.*, vol. 96, no. 1150, Oct. 2023, Art. no. 20221031.
[10] B. Koçak, A. Ponsiglione, A. Stanzione, C. Bluethgen, J. Santinha, L. Ugga, M. Huisman, M. E. Klontzas, R. Cannella, and R. Cuocolo, "Bias in artificial intelligence for medical imaging: Fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects," *Diagnostic Int. Radiol.*, vol. 31, no. 2, p. 75, Jul. 2024.
[11] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.
[12] N. Rane, S. Choudhary, and J. Rane, "Explainable artificial intelligence (XAI) in healthcare: Interpretable models for clinical decision support," *SSRN Electron. J.*, 2023.
[13] U. Kose, N. Sengoz, X. Chen, and J. A. M. Saucedo, *Explainable Artificial Intelligence (XAI) in Healthcare*. Boca Raton, FL, USA: CRC Press, 2024.
[14] N. Littlefield, H. Moradi, S. Amirian, H. M. Kremers, J. F. Plate, and A. P. Tafti, "Enforcing explainable deep few-shot learning to analyze plain knee radiographs: Data from the osteoarthritis initiative," in *Proc. IEEE 11th Int. Conf. Healthcare Informat. (ICHI)*, Jun. 2023, pp. 252–260.
[15] S. Amirian, Z. Wang, T. R. Taha, and H. R. Arabnia, "Dissection of deep learning with applications in image recognition," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Dec. 2018, pp. 1142–1148.
[16] M. Hassaballah and A. I. Awad, *Deep Learning in Computer Vision: Principles and Applications*. Boca Raton, FL, USA: CRC Press, 2020.
[17] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, no. 1, pp. 1–13, 2018.
[18] J. Chai, H. Zeng, A. Li, and E. W. T. Ngai, "Deep learning in computer vision: A critical review of emerging techniques and application scenarios," *Mach. Learn. Appl.*, vol. 6, Dec. 2021, Art. no. 100134.
[19] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102470.
[20] K. Borys, Y. A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C. M. Friedrich, and F. Nensa, "Explainable AI in medical imaging: An overview for clinical practitioners–beyond saliency-based XAI approaches," *Eur. J. Radiol.*, vol. 162, May 2023, Art. no. 110786.
[21] S. R. Sindiramutty, W. J. Tee, S. Balakrishnan, S. Kaur, R. Thangaveloo, H. Jazri, N. A. Khan, A. H. Gharib, and A. R. Manchuri, "Explainable AI in healthcare application," in *Advances in Explainable AI Applications for Smart Cities*. Hershey, PA, USA: IGI Global, 2024, pp. 123–176.

[22] S. Amirian, L. A. Carlson, M. F. Gong, I. Lohse, K. R. Weiss, J. F. Plate, and A. P. Tafti, "Explainable AI in orthopedics: Challenges, opportunities, and prospects," in *Proc. Congr. Comput. Sci., Comput. Eng., Appl. Comput. (CSCE)*, Jul. 2023, pp. 1374–1380.

[23] R. J. Chen, J. J. Wang, D. F. K. Williamson, T. Y. Chen, J. Lipkova, M. Y. Lu, S. Sahai, and F. Mahmood, "Algorithmic fairness in artificial intelligence for medicine and healthcare," *Nature Biomed. Eng.*, vol. 7, no. 6, pp. 719–742, Jun. 2023.

[24] D. Ueda, T. Kakinuma, S. Fujita, K. Kamagata, Y. Fushimi, R. Ito, Y. Matsui, T. Nozaki, T. Nakaura, N. Fujima, F. Tatsugami, M. Yanagawa, K. Hirata, A. Yamada, T. Tsuboyama, M. Kawamura, T. Fujioka, and S. Naganawa, "Fairness of artificial intelligence in healthcare: Review and recommendations," *Jpn. J. Radiol.*, vol. 42, no. 1, pp. 3–15, Jan. 2024.

[25] M. Liu, Y. Ning, S. Teixayavong, M. Mertens, J. Xu, D. S. W. Ting, L. T.-E. Cheng, J. C. L. Ong, Z. L. Teo, T. F. Tan, N. RaviChandran, F. Wang, L. A. Celi, M. E. H. Ong, and N. Liu, "A translational perspective towards clinical AI fairness," *NPJ Digit. Med.*, vol. 6, no. 1, p. 172, Sep. 2023.

[26] Centers Disease Control Prevention. (2024). *Why is Addressing SDOH Important?*. Accessed: Oct. 22, 2024.

[27] World Health Org. (2024). *Social Determinants of Health*. Accessed: Oct. 22, 2024.

[28] I. A. Siddiqui, N. Littlefield, L. A. Carlson, M. Gong, A. Chhabra, Z. Menezes, G. M. Mastorakos, S. M. Thakar, M. Abedian, I. Lohse, K. R. Weiss, J. F. Plate, H. Moradi, S. Amirian, and A. P. Tafti, "Fair AI-powered orthopedic image segmentation: Addressing bias and promoting equitable healthcare," *Sci. Rep.*, vol. 14, no. 1, p. 16105, Jul. 2024.

[29] Y. Yang, H. Zhang, J. W. Gichoya, D. Katabi, and M. Ghassemi, "The limits of fair medical imaging AI in real-world generalization," *Nature Med.*, vol. 30, no. 10, pp. 2838–2848, Oct. 2024.

[30] S. Tayebi Arasteh, A. Ziller, C. Kuhl, M. Makowski, S. Nebelung, R. Braren, D. Rueckert, D. Truhn, and G. Kaissis, "Preserving fairness and diagnostic accuracy in private large-scale AI models for medical imaging," *Commun. Med.*, vol. 4, no. 1, p. 46, Mar. 2024.

[31] M. A. Ricci Lara, R. Echeveste, and E. Ferrante, "Addressing fairness in artificial intelligence for medical imaging," *Nature Commun.*, vol. 13, no. 1, p. 4581, Aug. 2022.

[32] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, and N. Berthouze, "Evaluating saliency map explanations for convolutional neural networks: A user study," in *Proc. 25th Int. Conf. Intell. User Interfaces*, Mar. 2020, pp. 275–285.

[33] C.-Y. Hsu and W. Li, "Explainable GeoAI: Can saliency maps help interpret artificial intelligence's learning process? An empirical study on natural feature detection," *Int. J. Geographical Inf. Sci.*, vol. 37, no. 5, pp. 963–987, May 2023.

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[35] Y. Zhang, D. Hong, D. McClement, O. Oladosu, G. Pridham, and G. Slaney, "Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging," *J. Neurosci. Methods*, vol. 353, Apr. 2021, Art. no. 109098.

[36] L. Raatikainen and E. Rahtu, "The weighting game: Evaluating quality of explainability methods," 2022, *arXiv:2208.06175*.

[37] V. Pillai and H. Pirsiavash, "Explainable models with consistent interpretations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, May 2021, pp. 2431–2439.

[38] J. Edin, A. G. Motzfeldt, C. L. Christensen, T. Ruotsalo, L. Maaløe, and M. Maistro, "Normalized AOPC: Fixing misleading faithfulness metrics for feature attribution explainability," 2024, *arXiv:2408.08137*.

[39] Z. Yu, J. Chakraborty, and T. Menzies, "FairBalance: How to achieve equalized odds with data pre-processing," *IEEE Trans. Softw. Eng.*, vol. 50, no. 9, pp. 2294–2312, Sep. 2024.

[40] D. Muhammad and M. Bendechache, "Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis," *Comput. Struct. Biotechnol. J.*, vol. 24, pp. 542–560, Dec. 2024.

[41] K. H. Kim, H.-W. Koo, B.-J. Lee, S.-W. Yoon, and M.-J. Sohn, "Cerebral hemorrhage detection and localization with medical imaging for cerebrovascular disease diagnosis and treatment using explainable deep learning," *J. Korean Phys. Soc.*, vol. 79, no. 3, pp. 321–327, Aug. 2021.

[42] M. I. Hossain, G. Zamzmi, P. R. Mouton, M. S. Salekin, Y. Sun, and D. Goldgof, "Explainable AI for medical data: Current methods, limitations, and future directions," *ACM Comput. Surveys*, vol. 57, no. 6, pp. 1–46, Jun. 2025.

[43] S. K. Vuppala, M. Behera, H. Jack, and N. Bussa, "Explainable deep learning methods for medical imaging applications," in *Proc. IEEE 5th Int. Conf. Comput. Commun. Autom. (ICCCA)*, Oct. 2020, pp. 334–339.

[44] R. Correa, M. Shaan, H. Trivedi, B. Patel, L. A. G. Celi, J. W. Gichoya, and I. Banerjee, "A systematic review of 'fair' AI model development for image classification and prediction," *J. Med. Biol. Eng.*, vol. 42, no. 6, pp. 816–827, Dec. 2022.

[45] M. O. Khan, M. M. Afzal, S. Mirza, and Y. Fang, "How fair are medical imaging foundation models?" in *Proc. Mach. Learn. Health (ML4H)*, 2023, pp. 217–231.

[46] B. Allen, "The promise of explainable AI in digital health for precision medicine: A systematic review," *J. Personalized Med.*, vol. 14, no. 3, p. 277, Mar. 2024.

[47] J. Rane, Ö. Kaya, S. K. Mallick, and N. L. Rane, "Enhancing black-box models: Advances in explainable artificial intelligence for ethical decision-making," in *Future Research Opportunities for Artificial Intelligence in Industry*, vol. 5, Oct. 2024, p. 2.

[48] D. A. Tuan, "Bridging the gap between black box AI and clinical practice: Advancing explainable AI for trust, ethics, and personalized healthcare diagnostics," Tech. Rep., 2024.

[49] K. Lekadir, R. Osuala, C. Gallin, N. Lazrak, K. Kushibar, G. Tsakou, S. Aussó, L. C. Alberich, K. Marias, M. Tsiknakis, S. Colantonio, N. Papanikolaou, Z. Salahuddin, H. C. Woodruff, P. Lambin, and L. Martí-Bonmatí, "FUTURE-AI: Guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging," 2021, *arXiv:2109.09658*.

[50] D. Kollias, A. Arsenos, and S. Kollias, "Domain adaptation, explainability & fairness in AI for medical image analysis: Diagnosis of COVID-19 based on 3-D chest CT-scans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 4907–4914.

[51] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Comput. Methods Programs Biomed.*, vol. 226, Nov. 2022, Art. no. 107161.

[52] M. R. Karim, J. Jiao, T. Döhmen, M. Cochez, O. Beyan, D. Rebholz-Schuhmann, and S. Decker, "DeepKneeExplainer: Explainable knee osteoarthritis diagnosis from radiographs and magnetic resonance imaging," *IEEE Access*, vol. 9, pp. 39757–39780, 2021.

[53] S. Pereira, R. Meier, V. Alves, M. Reyes, and C. A. Silva, "Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment," in *Proc. 1st Int. Workshops Understand. Interpreting Mach. Learn. Med. Image Comput. Appl.*, Granada, Spain, Jan. 2018, pp. 106–114.

[54] M. Esmaeili, R. Vettukattil, H. Banitalebi, N. R. Krogh, and J. T. Geitung, "Explainable artificial intelligence for human–machine interaction in brain tumor localization," *J. Personalized Med.*, vol. 11, no. 11, p. 1213, Nov. 2021.

[55] B. Aldughayfiq, F. Ashfaq, N. Z. Jhanjhi, and M. Humayun, "Explainable AI for retinoblastoma diagnosis: Interpreting deep learning models with LIME and SHAP," *Diagnostics*, vol. 13, no. 11, p. 1932, Jun. 2023.

[56] F. Ursin, C. Timmermann, and F. Steger, "Explicability of artificial intelligence in radiology: Is a fifth bioethical principle conceptually necessary?" *Bioethics*, vol. 36, no. 2, pp. 143–153, Feb. 2022.

[57] B. H. M. van der Velden, "Explainable AI: Current status and future potential," *Eur. Radiol.*, vol. 34, no. 2, pp. 1187–1189, Aug. 2023.

[58] H. Saleem, A. R. Shahid, and B. Raza, "Visual interpretability in 3D brain tumor segmentation network," *Comput. Biol. Med.*, vol. 133, Jun. 2021, Art. no. 104410.

[59] E. Puyol-Antón, B. Ruijsink, S. K. Piechnik, S. Neubauer, S. E. Petersen, R. Razavi, and A. P. King, "Fairness in cardiac MR image analysis: An investigation of bias due to data imbalance in deep learning based segmentation," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, Strasbourg, France, Jan. 2021, pp. 413–423.

[60] K. Raghavan, S. Balasubramanian, and K. Veezhinathan, "Explainable artificial intelligence for medical imaging: Review and experiments with infrared breast images," *Comput. Intell.*, vol. 40, no. 3, Jun. 2024, Art. no. e12660.

[61] D. Mukhtorov, M. Rakhmonova, S. Muksimova, and Y.-I. Cho, "Endoscopic image classification based on explainable deep learning," *Sensors*, vol. 23, no. 6, p. 3176, Mar. 2023.
[62] S. M. Javali, R. Surya Upadhyayula, and T. De, "Comparative study of xAI layer-wise algorithms with a robust recommendation framework of inductive clustering for polyp segmentation and classification," in *Proc. Int. Seminar Mach. Learn., Optim., Data Sci. (ISMODE)*, Jan. 2022, pp. 325–330.
[63] S. M. Hussain, D. Buongiorno, N. Altini, F. Berloco, B. Prencipe, M. Moschetta, V. Bevilacqua, and A. Brunetti, "Shape-based breast lesion classification using digital tomosynthesis images: The role of explainable artificial intelligence," *Appl. Sci.*, vol. 12, no. 12, p. 6230, Jun. 2022.
[64] L. M. Duamwan and J. J. Bird, "Explainable AI for medical image processing: A study on MRI in Alzheimer's disease," in *Proc. 16th Int. Conf. Pervasive Technol. Rel. Assistive Environments*, Jul. 2023, pp. 480–484.
[65] H. Zhang, N. Dullerud, K. Roth, L. Oakden-Rayner, S. Pfohl, and M. Ghassemi, "Improving the fairness of chest X-ray classifiers," in *Proc. Conf. Health, Inference, Learn.*, vol. 174, G. Flores, G. H. Chen, T. Pollard, J. C. Ho, and T. Naumann, Eds., Jan. 2022, pp. 204–233.
[66] C. Patrício, J. C. Neves, and L. F. Teixeira, "Explainable deep learning methods in medical image classification: A survey," *ACM Comput. Surveys*, vol. 56, no. 4, pp. 1–41, Apr. 2024.
[67] Y. Tian, C. Wen, M. Shi, M. M. Afzal, H. Huang, M. O. Khan, Y. Luo, Y. Fang, and M. Wang, "FairDomain: Achieving fairness in cross-domain medical image segmentation and classification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2024, pp. 251–271.
[68] G. B. Bordes, "Fairness and explainability in chest X-ray image classifiers," M.S. thesis, Dept.Bioinformatics Biocomplexity, Utrecht Univ., Utrecht, The Netherlands, 2023.
[69] S. Ghosh, H. M. Abushukair, A. Ganesan, C. Pan, A. R. Naqash, and K. Lu, "Harnessing explainable artificial intelligence for patient-to-clinical-trial matching: A proof-of-concept pilot study using phase i oncology trials," *PLoS ONE*, vol. 19, no. 10, Oct. 2024, Art. no. e0311510.
[70] C.-Y. Chang, J. Yuan, S. Ding, Q. Tan, K. Zhang, X. Jiang, H. Xia, and N. Zou, "Towards fair patient-trial matching via patient-criterion level fairness constraint," in *Proc. AMIA Annu. Symp.*, Jan. 2023, p. 884.
[71] E. Ferrara, "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies," *Science*, vol. 6, no. 1, p. 3, Dec. 2023.
[72] A. S. Tejani, Y. S. Ng, Y. Xi, and J. C. Rayan, "Understanding and mitigating bias in imaging artificial intelligence," *RadioGraphics*, vol. 44, no. 5, May 2024, Art. no. e230067.
[73] D. Leben, "Explainable AI as evidence of fair decisions," *Frontiers Psychol.*, vol. 14, Feb. 2023, Art. no. 1069426.
[74] A. V. Menon, Z. A. Omar, N. Nahar, X. Papademetris, L. E. Fiellin, and C. Kästner, "Lessons from clinical communications for explainable AI," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, vol. 7, Oct. 2024, pp. 958–970.
[75] C. Novelli, M. Taddeo, and L. Floridi, "Accountability in artificial intelligence: What it is and how it works," *AI Soc.*, vol. 39, no. 4, pp. 1871–1882, Aug. 2024.
[76] M. Anagnostou, O. Karvounidou, C. Katritzidaki, C. Kechagia, K. Melidou, E. Mpeza, I. Konstantinidis, E. Kapantai, C. Berberidis, I. Magnisalis, and V. Peristeras, "Characteristics and challenges in the industries towards responsible AI: A systematic literature review," *Ethics Inf. Technol.*, vol. 24, no. 3, p. 37, Sep. 2022.
[77] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, K. Scott, S. Schieber, J. Waldo, D. Weinberger, A. Weller, and A. Wood, "Accountability of AI under the law: The role of explanation," 2017, *arXiv:1711.01134*.
[78] N. S. Germanov, "The concept of responsible artificial intelligence as the future of artificial intelligence in medicine," *Digit. Diag.*, vol. 4, no. 1S, pp. 27–29, Jun. 2023.
[79] K. Werder, B. Ramesh, and R. Zhang, "Establishing data provenance for responsible artificial intelligence systems," *ACM Trans. Manage. Inf. Syst.*, vol. 13, no. 2, pp. 1–23, Jun. 2022.
[80] I. Habli, T. Lawton, and Z. Porter, "Artificial intelligence in health care: Accountability and safety," *Bull. World Health Org.*, vol. 98, no. 4, pp. 251–256, Apr. 2020.
[81] A. M. Kempton and P. Vassilakopoulou, "Accountability, transparency and explainability in AI for healthcare," in *Proc. 8th Int. Conf. Infrastructures Healthcare*, Jan. 2021, pp. 1–10.
[82] N. Khalid, A. Qayyum, M. Bilal, A. Al-Fuqaha, and J. Qadir, "Privacy-preserving artificial intelligence in healthcare: Techniques and applications," *Comput. Biol. Med.*, vol. 158, May 2023, Art. no. 106848.
[83] M. Ali, F. Naeem, M. Tariq, and G. Kaddoum, "Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 2, pp. 778–789, Feb. 2023.
[84] S. M. Williamson and V. Prybutok, "Balancing privacy and progress: A review of privacy challenges, systemic oversight, and patient perceptions in AI-driven healthcare," *Appl. Sci.*, vol. 14, no. 2, p. 675, Jan. 2024.
[85] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nature Mach. Intell.*, vol. 2, no. 6, pp. 305–311, Jun. 2020.
[86] S. Selvakanmani, G. D. Devi, V. Rekha, and J. Jeyalakshmi, "Privacy-preserving breast cancer classification: A federated transfer learning approach," *J. Imag. Informat. Med.*, vol. 37, no. 4, pp. 1488–1504, Feb. 2024.
[87] E. Frimpong, K. Nguyen, M. Budzys, T. Khan, and A. Michalas, "GuardML: Efficient privacy-preserving machine learning services through hybrid homomorphic encryption," in *Proc. 39th ACM/SIGAPP Symp. Appl. Comput.*, Apr. 2024, pp. 953–962.
[88] A. Vizitiu, C. I. Nita, A. Puiu, C. Suciu, and L. M. Itu, "Towards privacy-preserving deep learning based medical imaging applications," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Jun. 2019, pp. 1–6.
[89] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, Jan. 2016, pp. 1273–1282.
[90] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons With Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)*, document OJL119, 2016.
[91] N. P. Terry, "Of regulating healthcare AI and robots," *Yale JL Tech.*, vol. 21, p. 133, Jul. 2019.
[92] J. Meszaros, J. Minari, and I. Huys, "The future regulation of artificial intelligence systems in healthcare services and medical research in the European union," *Frontiers Genet.*, vol. 13, Oct. 2022, Art. no. 927721.
[93] P. Keyur, "Lawful and righteous considerations for the use of artificial intelligence in public health," *Int. J. Comput. Trends Technol.*, vol. 72, no. 1, pp. 48–52, Jan. 2024.
[94] V. A. Laptev, I. V. Ershova, and D. R. Feyzrakhmanova, "Medical applications of artificial intelligence (legal aspects and future prospects)," *Laws*, vol. 11, no. 1, p. 3, Dec. 2021.
[95] S. Nasir, R. A. Khan, and S. Bai, "Ethical framework for harnessing the power of AI in healthcare and beyond," *IEEE Access*, vol. 12, pp. 31014–31035, 2024.
[96] J. A. Kroll, "Outlining traceability: A principle for operationalizing accountability in computing systems," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, Mar. 2021, pp. 758–771.
[97] M. Mora-Cantallops, S. Sánchez-Alonso, E. García-Barriocanal, and M.-A. Sicilia, "Traceability for trustworthy AI: A review of models and tools," *Big Data Cogn. Comput.*, vol. 5, no. 2, p. 20, May 2021.
[98] R. Souza, L. Azevedo, V. Lourenço, E. Soares, R. Thiago, R. Brandão, D. Civitarese, E. Brazil, M. Moreno, P. Valduriez, M. Mattoso, R. Cerqueira, and M. A. S. Netto, "Provenance data in the machine learning lifecycle in computational science and engineering," in *Proc. IEEE/ACM Workflows Support Large-Scale Sci. (WORKS)*, Nov. 2019, pp. 1–10.
[99] P. Gkontra, G. Quaglio, A. T. Garmendia, and K. Lekadir, "Challenges of machine learning and AI (what is next?), responsible and ethical AI," in *Clinical Applications of Artificial Intelligence in Real-World Data*. Cham, Switzerland: Springer, 2023, pp. 263–285.
[100] A. Chakraborty and M. Karhade, "Global AI governance in healthcare: A cross-jurisdictional regulatory analysis," 2024, *arXiv:2406.08695*.
[101] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, "Trustworthy AI: From principles to practices," *ACM Comput. Surveys*, vol. 55, no. 9, pp. 1–46, Aug. 2022.
[102] J. Mahilraj, M. Pandian, M. Subbiah, S. Kalyan, R. Vadivel, and S. Nirmala, "Evaluation of the robustness, transparency, reliability and safety of AI systems," in *Proc. 9th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, vol. 1, Mar. 2023, pp. 2526–2535.

[103] E. Petersen, Y. Potdevin, E. Mohammadi, S. Zidowitz, S. Breyer, D. Nowotka, S. Henn, L. Pechmann, M. Leucker, P. Rostalski, and C. Herzog, "Responsible and regulatory conform machine learning for medicine: A survey of challenges and solutions," *IEEE Access*, vol. 10, pp. 58375–58418, 2022.

[104] Y. Balagurunathan, R. Mitchell, and I. El Naqa, "Requirements and reliability of AI in the medical context," *Phys. Medica*, vol. 83, pp. 72–78, Mar. 2021.

[105] O. Higgins, S. K. Chalup, and R. L. Wilson, "Artificial intelligence in nursing: Trustworthy or reliable?" *J. Res. Nursing*, vol. 29, no. 2, pp. 143–153, Mar. 2024.

[106] S. Polevikov, "Advancing AI in healthcare: A comprehensive review of best practices," *Clinica Chim. Acta*, vol. 548, Aug. 2023, Art. no. 117519.

[107] N. M. Thomasian, C. Eickhoff, and E. Y. Adashi, "Advancing health equity with artificial intelligence," *J. Public Health Policy*, vol. 42, no. 4, pp. 602–611, Dec. 2021.

[108] M. A. Davis, N. Lim, J. Jordan, J. Yee, J. W. Gichoya, and R. Lee, "Imaging artificial intelligence: A framework for radiologists to address health equity, from the AJR special series on DEI," *Amer. J. Roentgenology*, vol. 221, no. 3, pp. 302–308, Sep. 2023.

[109] D. Ng, X. Lan, M. M.-S. Yao, W. P. Chan, and M. Feng, "Federated learning: A collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets," *Quant. Imag. Med. Surg.*, vol. 11, no. 2, pp. 852–857, Feb. 2021.

[110] L. Kwak and H. Bai, "The role of federated learning models in medical imaging," *Radiol., Artif. Intell.*, vol. 5, no. 3, May 2023, Art. no. e230136.

[111] S. Reddy, S. Allan, S. Coghlan, and P. Cooper, "A governance model for the application of AI in health care," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 3, pp. 491–497, Mar. 2020.

[112] Y. Y. M. Aung, D. C. S. Wong, and D. S. W. Ting, "The promise of artificial intelligence: A review of the opportunities and challenges of artificial intelligence in healthcare," *Brit. Med. Bull.*, vol. 139, no. 1, pp. 4–15, Sep. 2021.

[113] U. K. Kar, "The future of health and healthcare in a world of artificial intelligence," *Arch. Biomed. Eng. Biotechnol.*, vol. 1, no. 1, pp. 1–7, 2018.

[114] M. Hassan, E. M. Borycki, and A. W. Kushniruk, "Artificial intelligence governance framework for healthcare," *Healthcare Manage. Forum*, vol. 38, no. 2, pp. 125–130, Mar. 2025.

[115] G. B. Mensah, F. Nyante, A. Addy, and P. O. Frimpong, "The legal, ethical, and regulatory implications of integrating artificial intelligence in healthcare delivery, medical negligence, and public health administration in Ghana: A multidisciplinary perspective," *Afr. J. Regulatory Affairs*, vol. 2024, pp. 1–13, Jul. 2024.

[116] J. K. Wagner, M. Doerr, and C. D. Schmit, "AI governance: A challenge for public health," *JMIR Public Health Surveill.*, vol. 10, no. 1, Sep. 2024, Art. no. e58358.

[117] M. Hassan, A. Kushniruk, and E. Borycki, "Barriers to and facilitators of artificial intelligence adoption in health care: Scoping review," *JMIR Human Factors*, vol. 11, Aug. 2024, Art. no. e48633.

[118] A. Singla and T. Malhotra, "Challenges and opportunities in scaling AI/ML pipelines," *J. Sci. Technol.*, vol. 5, no. 1, pp. 1–21, 2024.

[119] H. Barmer, R. Dzombak, M. Gaston, V. Palat, F. Redner, T. Smith, and J. Wohlbier, "Scalable AI," Tech. Rep., 2021.

[120] R. Y. Cohen and V. P. Kovacheva, "A methodology for a scalable, collaborative, and resource-efficient platform to facilitate healthcare AI research," 2021, *arXiv:2112.06883*.

[121] K. Walia, "Scalable AI models through cloud infrastructure," *ESP Int. J. Advancements Comput. Technol.*, vol. 2, no. 2, pp. 1–7, 2024.

[122] W. Brewer, A. Kashi, S. Dash, A. Tsaris, J. Yin, M. Shankar, and F. Wang, "Scalable artificial intelligence for science: Perspectives, methods and exemplars," 2024, *arXiv:2406.17812*.

[123] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowl.-Based Syst.*, vol. 216, Jan. 2021, Art. no. 106775.

[124] P. Pinyoanuntapong, W. H. Huff, M. Lee, C. Chen, and P. Wang, "Toward scalable and robust AIoT via decentralized federated learning," *IEEE Internet Things Mag.*, vol. 5, no. 1, pp. 30–35, Mar. 2022.

[125] T. T. Nguyen, T. Silander, Z. Li, and T.-Y. Leong, "Scalable transfer learning in heterogeneous, dynamic environments," *Artif. Intell.*, vol. 247, pp. 70–94, Jun. 2017.

[126] A. Bohr and K. Memarzadeh, "The rise of artificial intelligence in healthcare applications," in *Artificial Intelligence in Healthcare*. Amsterdam, The Netherlands: Elsevier, 2020, pp. 25–60.

[127] A. Panesar, *Machine Learning and AI for Healthcare*. Cham, Switzerland: Springer, 2019.

[128] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future Healthcare J.*, vol. 6, no. 2, pp. 94–98, Jun. 2019.

[129] A. A. Kuwaiti, K. Nazer, A. H. Alreedy, S. D. AlShehri, A. Almuhanna, A. V. Subbarayalu, D. A. Muhanna, and F. Al-Muhanna, "A review of the role of artificial intelligence in healthcare," *J. Personalized Med.*, vol. 13, no. 6, p. 951, Jun. 2023.

[130] A. A. Shick, C. M. Webber, N. Kiarashi, J. P. Weinberg, A. Deoras, N. Petrick, A. Saha, and M. C. Diamond, "Transparency of artificial intelligence/machine learning-enabled medical devices," *NPJ Digit. Med.*, vol. 7, no. 1, p. 21, Jan. 2024.

[131] Z. Sadeghi, R. Alizadehsani, M. A. Çifçi, S. Kausar, R. Rehman, P. Mahanta, P. K. Bora, A. Almasri, R. S. Alkhawaldeh, S. Hussain, B. Alataş, A. Shoeibi, H. Moosaei, M. Hladík, S. Nahavandi, and P. M. Pardalos, "A review of explainable artificial intelligence in healthcare," *Comput. Electr. Eng.*, vol. 118, Jun. 2024, Art. no. 109370.

[132] J. Fehr, B. Citro, R. Malpani, C. Lippert, and V. I. Madai, "A trustworthy AI reality-check: The lack of transparency of artificial intelligence products in healthcare," *Frontiers Digit. Health*, vol. 6, Feb. 2024, Art. no. 1267290.

[133] C. Metta, A. Beretta, R. Pellungrini, S. Rinzivillo, and F. Giannotti, "Towards transparent healthcare: Advancing local explanation methods in explainable artificial intelligence," *Bioengineering*, vol. 11, no. 4, p. 369, Apr. 2024.

[134] F. C. Oettl, J. F. Oeding, and K. Samuelsson, "Explainable artificial intelligence in orthopedic surgery," *J. Exp. Orthopaedics*, vol. 11, no. 3, Jul. 2024, Art. no. e12103.

[135] J. Amann, D. Vetter, S. N. Blomberg, H. C. Christensen, M. Coffee, S. Gerke, T. K. Gilbert, T. Hagendorff, S. Holm, M. Livne, A. Spezzatti, I. Strümke, R. V. Zicari, and V. I. Madai, "To explain or not to explain—Artificial intelligence explainability in clinical decision support systems," *PLOS Digit. Health*, vol. 1, no. 2, Feb. 2022, Art. no. e0000016.

[136] D. K. Chettri, "Explainable AI for decision-making systems: Investigate the development of explainable AI techniques for decision-making systems and evaluate their effectiveness in improving the transparency and accountability of these systems," *Int. J. Mod. Develop. Eng. Sci.*, vol. 2, no. 2, pp. 1–6, 2023.

[137] G. A. S. Thomas, S. Muthukaruppasamy, J. N. Gopal, G. Sudha, and K. Saravanan, "Unleashing the power of XAI (explainable artificial intelligence): Empowering decision-making and overcoming challenges in smart healthcare automation," in *Explainable AI (XAI) for Sustainable Development: Trends and Applications*. Boca Raton, FL, USA: CRC Press, 2024, pp. 303–316.

[138] E. Häikiö and J. Heikkilä, "Adoption and governance of AI-powered dashboards in executive-level decision-making," Tech. Rep., 2024.

[139] A. Marey, P. Arjmand, A. D. S. Alerab, M. J. Eslami, A. M. Saad, N. Sanchez, and M. Umair, "Explainability, transparency and black box challenges of AI in radiology: Impact on patient care in cardiovascular radiology," *Egyptian J. Radiol. Nucl. Med.*, vol. 55, no. 1, pp. 1–14, Sep. 2024.

[140] H. Chen, C. Gomez, C.-M. Huang, and M. Unberath, "Explainable medical imaging AI needs human-centered design: Guidelines and evidence from a systematic review," *NPJ Digit. Med.*, vol. 5, no. 1, p. 156, Oct. 2022.

[141] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *J. Imag.*, vol. 6, no. 6, p. 52, Jun. 2020.

[142] M. Champendal, H. Müller, J. O. Prior, and C. S. Dos Reis, "A scoping review of interpretability and explainability concerning artificial intelligence methods in medical imaging," *Eur. J. Radiol.*, vol. 169, Dec. 2023, Art. no. 111159.

[143] M. Fontes, J. D. S. De Almeida, and A. Cunha, "Application of example-based explainable artificial intelligence (XAI) for analysis and interpretation of medical imaging: A systematic review," *IEEE Access*, vol. 12, pp. 26419–26427, 2024.

[144] D. Saraswat, P. Bhattacharya, A. Verma, V. K. Prasad, S. Tanwar, G. Sharma, P. N. Bokoro, and R. Sharma, "Explainable AI for Healthcare 5.0: Opportunities and challenges," *IEEE Access*, vol. 10, pp. 84486–84517, 2022.

[145] M. Aasem and M. Javed Iqbal, "Toward explainable AI in radiology: Ensemble-CAM for effective thoracic disease localization in chest X-ray images using weak supervised learning," *Frontiers Big Data*, vol. 7, May 2024, Art. no. 1366415.
[146] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, vol. 11700. Cham, Switzerland: Springer, 2019.
[147] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?" 2017, *arXiv:1712.09923*.
[148] A. P. Tafti, F. S. Bashiri, E. LaRose, and P. Peissig, "Diagnostic classification of lung CT images using deep 3D multi-scale convolutional neural network," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Jun. 2018, pp. 412–414.
[149] D. Blanc et al., "Artificial intelligence solution to classify pulmonary nodules on CT," *Diagnostic Interventional Imag.*, vol. 101, no. 12, pp. 803–810, Dec. 2020.
[150] R. Mehrotra, M. A. Ansari, R. Agrawal, and R. S. Anand, "A transfer learning approach for AI-based classification of brain tumors," *Mach. Learn. Appl.*, vol. 2, Dec. 2020, Art. no. 100003.
[151] G. S. Tandel, A. Balestrieri, T. Jujaray, N. N. Khanna, L. Saba, and J. S. Suri, "Multiclass magnetic resonance imaging brain tumor classification using artificial intelligence paradigm," *Comput. Biol. Med.*, vol. 122, Jul. 2020, Art. no. 103804.
[152] R. Ranjbarzadeh, A. Caputo, E. B. Tirkolaee, S. J. Ghoushchi, and M. Bendechache, "Brain tumor segmentation of MRI images: A comprehensive review on the application of artificial intelligence tools," *Comput. Biol. Med.*, vol. 152, Jan. 2023, Art. no. 106405.
[153] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies," *J. Biomed. Informat.*, vol. 113, Jan. 2021, Art. no. 103655.
[154] A. S. Albahri, A. M. Duhaim, M. A. Fadhel, A. Alnoor, N. S. Baqer, L. Alzubaidi, O. S. Albahri, A. H. Alamoodi, J. Bai, A. Salhi, J. Santamaría, C. Ouyang, A. Gupta, Y. Gu, and M. Deveci, "A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion," *Inf. Fusion*, vol. 96, pp. 156–191, Aug. 2023.
[155] K. Kostick-Quenet, B. H. Lang, J. Smith, M. Hurley, and J. Blumenthal-Barby, "Trust criteria for artificial intelligence in health: Normative and epistemic considerations," *J. Med. Ethics*, vol. 50, no. 8, pp. 544–551, Aug. 2024.
[156] M. Zandehshahvar, M. van Assen, E. Kim, Y. Kiarashi, V. Keerthipati, G. Tessarin, E. Muscogiuri, A. E. Stillman, P. Filev, A. H. Davarpanah, E. A. Berkowitz, S. Tigges, S. J. Lee, B. L. Vey, C. De Cecco, and A. Adibi, "Confidence-aware severity assessment of lung disease from chest X-rays using deep neural network on a multi-reader dataset," *J. Imag. Informat. Med.*, vol. 2024, pp. 1–11, Aug. 2024.
[157] S. H. A. Mahmood, Z. Lu, and M. Yin, "Designing behavior-aware AI to improve the human-AI team performance in AI-assisted decision making," in *Proc. 33rd Int. Joint Conf. Artif. Intell.*, Aug. 2024, pp. 3106–3114.
[158] C. Macrae, "Governing the safety of artificial intelligence in healthcare," *BMJ Quality Saf.*, vol. 28, no. 6, pp. 495–498, Jun. 2019.
[159] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, "Addressing bias in big data and AI for health care: A call for open science," *Patterns*, vol. 2, no. 10, Oct. 2021, Art. no. 100347.
[160] S. L. Sargent, "AI bias in healthcare: Using ImpactPro as a case study for healthcare practitioners' duties to engage in anti-bias measures," *Can. J. Bioethics*, vol. 4, no. 1, pp. 112–116, May 2021.
[161] G. B. Mensah, "Artificial intelligence and ethics: A comprehensive review of bias mitigation, transparency, and accountability in AI systems," *Preprint*, vol. 10, no. 1, pp. 1–26, Nov. 2023.
[162] R. Schwartz, R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, "Towards a standard for identifying and managing bias in artificial intelligence," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep., 2022, vol. 3.
[163] M. P. Cary, S. Bessias, J. McCall, M. J. Pencina, S. D. Grady, K. Lytle, and N. J. Economou-Zavlanos, "Empowering nurses to champion health equity & BE FAIR: Bias elimination for fair and responsible AI in healthcare," *J. Nursing Scholarship*, vol. 57, no. 1, pp. 130–139, Jan. 2025.
[164] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. L. De Prado, E. Herrera-Viedma, and F. Herrera, "Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101896.
[165] L. Szabo, Z. Raisi-Estabragh, A. Salih, C. McCracken, E. Ruiz Pujadas, P. Gkontra, M. Kiss, P. Maurovich-Horvath, H. Vago, B. Merkely, A. M. Lee, K. Lekadir, and S. E. Petersen, "Clinician's guide to trustworthy and responsible artificial intelligence in cardiovascular imaging," *Frontiers Cardiovascular Med.*, vol. 9, Nov. 2022, Art. no. 1016032.
[166] F. Li, N. Ruijs, and Y. Lu, "Ethics & AI: A systematic review on ethical concerns and related strategies for designing with AI in healthcare," *AI*, vol. 4, no. 1, pp. 28–53, Dec. 2022.
[167] D. Peters, K. Vold, D. Robinson, and R. A. Calvo, "Responsible AI—Two frameworks for ethical design practice," *IEEE Trans. Technol. Soc.*, vol. 1, no. 1, pp. 34–47, Mar. 2020.
[168] N. Vallès-Peris and M. Domènech, "Caring in the in-between: A proposal to introduce responsible AI and robotics to healthcare," *AI Soc.*, vol. 38, no. 4, pp. 1685–1695, Aug. 2023.
[169] M. Bekbolatova, J. Mayer, C. W. Ong, and M. Toma, "Transformative potential of AI in healthcare: Definitions, applications, and navigating the ethical landscape and public perspectives," *Healthcare*, vol. 12, no. 2, p. 125, Jan. 2024.
[170] B. Xia, Q. Lu, L. Zhu, S. U. Lee, Y. Liu, and Z. Xing, "Towards a responsible AI metrics catalogue: A collection of metrics for AI accountability," in *Proc. IEEE/ACM 3rd Int. Conf. AI Eng.-Softw. Eng. AI*, Apr. 2024, pp. 100–111.
[171] Y. Li and S. Goel, "Artificial intelligence auditability and auditor readiness for auditing artificial intelligence systems," *Int. J. Accounting Inf. Syst.*, vol. 56, Dec. 2025, Art. no. 100739.
[172] W. Murikah, J. K. Nthenge, and F. M. Musyoka, "Bias and ethics of AI systems applied in auditing—A systematic review," *Sci. Afr.*, vol. 25, Sep. 2024, Art. no. e02281.
[173] S. Toktas, "Auditing of AI," Tech. Rep., 2024.
[174] P. Esmaeilzadeh, "Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: A perspective for healthcare organizations," *Artif. Intell. Med.*, vol. 151, May 2024, Art. no. 102861.
[175] M. A. Camilleri, "Artificial intelligence governance: Ethical considerations and implications for social responsibility," *Expert Syst.*, vol. 41, no. 7, Jul. 2024, Art. no. e13406.
[176] Q. Lu, L. Zhu, X. Xu, J. Whittle, D. Zowghi, and A. Jacquet, "Responsible AI pattern catalogue: A collection of best practices for AI governance and engineering," *ACM Comput. Surveys*, vol. 56, no. 7, pp. 1–35, Jul. 2024.
[177] D. V. Ligot, "AI governance: A framework for responsible AI development," *SSRN Electron. J.*, 2024.
[178] C. W. L. Ho, D. Soon, K. Caals, and J. Kapur, "Governance of automated image analysis and artificial intelligence analytics in healthcare," *Clin. Radiol.*, vol. 74, no. 5, pp. 329–337, May 2019.
[179] S. Khanna and S. Srivastava, "AI governance in healthcare: Explainability standards, safety protocols, and human-AI interactions dynamics in contemporary medical AI systems," *Empirical Quests Manage. Essences*, vol. 1, no. 1, pp. 130–143, 2021.
[180] J. Guan, "Artificial intelligence in healthcare and medicine: Promises, ethical challenges, and governance," *Chin. Med. Sci. J.*, vol. 34, no. 2, pp. 76–83, Jan. 2019.
[181] U.S. Dept. Health Human Services. (1996). *Health Insurance Portability and Accountability Act of 1996*. Accessed: Nov. 19, 2024.
[182] B. Wolford. (2018). *What is GDPR, the EU's New Data Protection Law?*. Accessed: Nov. 20, 2024.
[183] J. Morley, L. Murphy, A. Mishra, I. Joshi, and K. Karpathakis, "Governing data and artificial intelligence for health care: Developing an international understanding," *JMIR Formative Res.*, vol. 6, no. 1, Jan. 2022, Art. no. e31623.
[184] Q. Yang, "Toward responsible AI: An overview of federated learning for user-centered privacy-preserving computing," *ACM Trans. Interact. Intell. Syst.*, vol. 11, nos. 3–4, pp. 1–22, Dec. 2021.
[185] A. C. Yao, "Protocols for secure computations," in *Proc. 23rd Annu. Symp. Found. Comput. Sci. (SFCS)*, Nov. 1982, pp. 160–164.

[186] C. Dwork, "Differential privacy: A survey of results," in *Proc. Int. Conf. Theory Appl. Models Comput.*, Apr. 2008, pp. 1–19.
[187] T. Garfinkel, B. Pfaff, J. Chow, M. Rosenblum, and D. Boneh, "Terra: A virtual machine-based platform for trusted computing," in *Proc. 19th ACM Symp. Operating Syst. Princ.-SOSP*, 2003, pp. 193–206.
[188] J. Rudd and C. Igbrude, "A global perspective on data powering responsible AI solutions in health applications," *AI Ethics*, vol. 4, no. 4, pp. 1039–1049, Nov. 2024.
[189] M. E. Hurley, B. H. Lang, K. M. Kostick-Quenet, J. N. Smith, and J. Blumenthal-Barby, "Patient consent and the right to notice and explanation of AI systems used in health care," *Amer. J. Bioethics*, vol. 25, no. 3, pp. 102–114, Mar. 2025.
[190] A. L. Kotsenas, P. Balthazar, D. Andrews, J. R. Geis, and T. S. Cook, "Rethinking patient consent in the era of artificial intelligence and big data," *J. Amer. College Radiol.*, vol. 18, no. 1, pp. 180–184, Jan. 2021.
[191] H. Torkamaan, M. Tahaei, S. Buijsman, Z. Xiao, D. Wilkinson, and B. P. Knijnenburg, "The role of human-centered AI in user modeling, adaptation, and personalization—Models, frameworks, and paradigms," in *A Human-Centered Perspective of Intelligent Personalized Environments and Systems*. Cham, Switzerland: Springer, 2024, pp. 43–83.
[192] V. Dignum, *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, vol. 2156. Cham, Switzerland: Springer, 2019.
[193] M. Nevanperä, "Aspects to responsible artificial intelligence: Ethics of artificial intelligence and ethical guidelines in shapes project," M.S. thesis, Laurea Univ. Appl. Sci., Vantaa, Finland, 2021.
[194] L. Oberste and A. Heinzl, "User-centric explainability in healthcare: A knowledge-level perspective of informed machine learning," *IEEE Trans. Artif. Intell.*, vol. 4, no. 4, pp. 840–857, Apr. 2022.
[195] A. H. Hassan, R. B. Sulaiman, M. A. Abdulgabber, and H. Kahtan, "Balancing technological advances with user needs: User-centered principles for AI-driven smart city healthcare monitoring," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 3, pp. 1–12, 2023.
[196] M. F. Almufareh, S. Kausar, M. Humayun, and S. Tehsin, "A conceptual model for inclusive technology: Advancing disability inclusion through artificial intelligence," *J. Disability Res.*, vol. 3, no. 1, Jan. 2024, Art. no. 20230060.
[197] B. A. Jnr., "User-centered AI-based voice-assistants for safe mobility of older people in urban context," *AI Soc.*, vol. 2024, pp. 1–24, Feb. 2024.
[198] C. El Morr, B. Kundi, F. Mobeen, S. Taleghani, Y. El-Lahib, and R. Gorman, "AI and disability: A systematic scoping review," *Health Informat. J.*, vol. 30, no. 3, Jul. 2024, Art. no. 14604582241285743.
[199] C.-Y. Wang and F.-S. Lin, "AI-driven privacy in elderly care: Developing a comprehensive solution for camera-based monitoring of older adults," *Appl. Sci.*, vol. 14, no. 10, p. 4150, May 2024.
[200] P. Radanliev, O. Santos, A. Brandon-Jones, and A. Joinson, "Ethics and responsible AI deployment," *Frontiers Artif. Intell.*, vol. 7, Mar. 2024, Art. no. 1377011.
[201] C. Sanderson, Q. Lu, D. Douglas, X. Xu, L. Zhu, and J. Whittle, "Towards implementing responsible AI," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2022, pp. 5076–5081.
[202] Q. Lu, L. Zhu, X. Xu, J. Whittle, D. Douglas, and C. Sanderson, "Software engineering for responsible AI: An empirical study and operationalised patterns," in *Proc. IEEE/ACM 44th Int. Conf. Softw. Eng., Softw. Eng. Pract. (ICSE-SEIP)*, May 2022, pp. 241–242.
[203] S. Nurhaliza and A. Van Nguyen, "Ethical AI governance: Principles, policies, and practices for responsible artificial intelligence," *Int. J. Appl. Health Care Anal.*, vol. 5, no. 12, pp. 28–36, 2020.
[204] Y. Wang, M. Xiong, and H. Olya, "Toward an understanding of responsible artificial intelligence practices," in *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, 2020, pp. 4962–4971.
[205] D. Domínguez Figaredo and J. Stoyanovich, "Responsible AI literacy: A stakeholder-first approach," *Big Data Soc.*, vol. 10, no. 2, Jul. 2023, Art. no. 20539517231219958.
[206] N. Economou, "Principles for the trustworthy adoption of AI in legal systems: The IEEE global initiative on ethics of autonomous and intelligent systems," in *Proc. LegalAIIA@ ICAIL*, Jan. 2019, pp. 2–5.
[207] R. Chatila, K. Firth-Butterflied, J. C. Havens, and K. Karachalios, "The IEEE global initiative for ethical considerations in artificial intelligence and autonomous systems [standards]," *IEEE Robot. Autom. Mag.*, vol. 24, no. 1, p. 110, Mar. 2017.
[208] X. Bai, X. Wang, X. Liu, Q. Liu, J. Song, N. Sebe, and B. Kim, "Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments," *Pattern Recognit.*, vol. 120, Dec. 2021, Art. no. 108102.
[209] K. Abhishek and D. Kamath, "Attribution-based XAI methods in computer vision: A review," 2022, *arXiv:2211.14736*.
[210] A. Perotti, C. Borile, A. Miola, F. P. Nerini, P. Baracco, and A. Panisson, "Explainability, quantified: Benchmarking XAI techniques," in *Proc. World Conf. Explainable Artif. Intell.*, Arianna Miola, France. Cham, Switzerland: Springer, Jan. 2024, pp. 421–444.
[211] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
[212] M. Miró-Nicolau, G. Moyá-Alcover, and A. Jaume-I-Capó, "Evaluating explainable artificial intelligence for X-ray image analysis," *Appl. Sci.*, vol. 12, no. 9, p. 4459, Apr. 2022.
[213] J. Wang, A. Bhalerao, T. Yin, S. See, and Y. He, "CAMANet: Class activation map guided attention network for radiology report generation," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 4, pp. 2199–2210, Apr. 2024.
[214] H. Jung and Y. Oh, "Towards better explanations of class activation mapping," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1316–1324.
[215] S. Shinde, T. Chougule, J. Saini, and M. Ingalhalikar, "HR-CAM: Precise localization of pathology using multi-level learning in CNNs," in *Proc. 22nd Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, Shenzhen, China. Cham, Switzerland: Springer, Jan. 2019, pp. 298–306.
[216] K. Gao, H. Shen, Y. Liu, L. Zeng, and D. Hu, "Dense-CAM: Visualize the gender of brains with MRI images," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–7.
[217] H. Lee, S. Yune, M. Mansouri, M. Kim, S. H. Tajmir, C. E. Guerrier, S. A. Ebert, S. R. Pomerantz, J. M. Romero, S. Kamalian, R. G. Gonzalez, M. H. Lev, and S. Do, "An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets," *Nature Biomed. Eng.*, vol. 3, no. 3, pp. 173–182, Dec. 2018.
[218] J. H. Lee, E. J. Ha, and J. H. Kim, "Application of deep learning to the diagnosis of cervical lymph node metastasis from thyroid cancer with CT," *Eur. Radiol.*, vol. 29, no. 10, pp. 5452–5457, Oct. 2019.
[219] Y. Lei, Y. Tian, H. Shan, J. Zhang, G. Wang, and M. K. Kalra, "Shape and margin-aware lung nodule classification in low-dose CT images via soft activation mapping," *Med. Image Anal.*, vol. 60, Feb. 2020, Art. no. 101628.
[220] L. Luo, H. Chen, X. Wang, Q. Dou, H. Lin, J. Zhou, G. Li, and P. Heng, "Deep angular embedding and feature correlation attention for breast MRI cancer analysis," in *Proc. 22nd Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, Shenzhen, China. Cham, Switzerland: Springer, Jan. 2019, pp. 504–512.
[221] L. Wang, L. Zhang, M. Zhu, X. Qi, and Z. Yi, "Automatic diagnosis for thyroid nodules in ultrasound images by deep neural networks," *Med. Image Anal.*, vol. 61, Apr. 2020, Art. no. 101665.
[222] W. Ausawalaithong, A. Thirach, S. Marukatat, and T. Wilaiprasitporn, "Automatic lung cancer prediction from chest X-ray images using the deep learning approach," in *Proc. 11th Biomed. Eng. Int. Conf. (BMEiCON)*, Nov. 2018, pp. 1–5.
[223] J. A. Dunnmon, D. Yi, C. P. Langlotz, C. Ré, D. L. Rubin, and M. P. Lungren, "Assessment of convolutional neural networks for automated classification of chest radiographs," *Radiology*, vol. 290, no. 2, pp. 537–544, Feb. 2019.
[224] W. Li, J. Zhuang, R. Wang, J. Zhang, and W.-S. Zheng, "Fusing metadata and dermoscopy images for skin disease diagnosis," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 1996–2000.
[225] P. Rajpurkar et al., "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLOS Med.*, vol. 15, no. 11, Nov. 2018, Art. no. e1002686.
[226] T. Lai, "Interpretable medical imagery diagnosis with self-attentive transformers: A review of explainable AI for health care," *BioMedInformatics*, vol. 4, no. 1, pp. 113–126, Jan. 2024.
[227] Y. Yang, G. Mei, and F. Piccialli, "A deep learning approach considering image background for pneumonia identification using explainable AI (XAI)," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 21, no. 4, pp. 857–868, Aug. 2024.
[228] D. D. Gunashekar, L. Bielak, L. Hägele, B. Oerther, M. Benndorf, A.-L. Grosu, T. Brox, C. Zamboglou, and M. Bock, "Explainable AI for CNN-based prostate tumor segmentation in multi-parametric MRI correlated to whole mount histopathology," *Radiat. Oncol.*, vol. 17, no. 1, p. 65, Dec. 2022.

[229] D. Varam, R. Mitra, M. Mkadmi, R. A. Riyas, D. A. Abuhani, S. Dhou, and A. Alzaatreh, "Wireless capsule endoscopy image classification: An explainable AI approach," *IEEE Access*, vol. 11, pp. 105262–105280, 2023.
[230] B. Korbar, A. M. Olofson, A. P. Miraflor, C. M. Nicka, M. A. Suriawinata, L. Torresani, A. A. Suriawinata, and S. Hassanpour, "Looking under the hood: Deep neural network visualization to interpret whole-slide image analysis outcomes for colorectal polyps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 821–827.
[231] P. Windisch, P. Weber, C. Fürweger, F. Ehret, M. Kufeld, D. Zwahlen, and A. Muacevic, "Implementation of model explainability for a basic brain tumor detection using convolutional neural networks on MRI slices," *Neuroradiology*, vol. 62, no. 11, pp. 1515–1518, Nov. 2020.
[232] K. A. Thakoor, X. Li, E. Tsamis, P. Sajda, and D. C. Hood, "Enhancing the accuracy of glaucoma detection from OCT probability maps using convolutional neural networks," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 2036–2040.
[233] T.-C. Lin and H.-C. Lee, "COVID-19 chest radiography images analysis based on integration of image preprocess, guided grad-CAM, machine learning and risk management," in *Proc. 4th Int. Conf. Med. Health Informat.*, Aug. 2020, pp. 281–288.
[234] A. Kajala, S. Jaiswal, and R. Kumar, "Breaking the black box: Heatmap-driven transparency to breast cancer detection with EfficientNet and Grad CAM," *Educ. Admin., Theory Pract.*, vol. 30, no. 5, pp. 4999–5009, 2024.
[235] F. M. Talaat, S. A. Gamel, R. M. El-Balka, M. Shehata, and H. ZainEldin, "Grad-CAM enabled breast cancer classification with a 3D inception-ResNet v2: Empowering radiologists with explainable insights," *Cancers*, vol. 16, no. 21, p. 3668, Oct. 2024.
[236] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2014, pp. 818–833.
[237] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.
[238] E. A. M. Stanley, M. Wilms, P. Mouches, and N. D. Forkert, "Fairness-related performance and explainability effects in deep learning models for brain image analysis," *J. Med. Imag.*, vol. 9, no. 6, Aug. 2022, Art. no. 061102.
[239] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.
[240] K. Yoon, J.-Y. Kim, S.-J. Kim, J.-K. Huh, J.-W. Kim, and J. Choi, "Explainable deep learning-based clinical decision support engine for MRI-based automated diagnosis of temporomandibular joint anterior disk displacement," *Comput. Methods Programs Biomed.*, vol. 233, May 2023, Art. no. 107465.
[241] H. Shin, J. E. Park, Y. Jun, T. Eo, J. Lee, J. E. Kim, D. H. Lee, H. H. Moon, S. I. Park, S. Kim, D. Hwang, and H. S. Kim, "Deep learning referral suggestion and tumour discrimination using explainable artificial intelligence applied to multiparametric MRI," *Eur. Radiol.*, vol. 33, no. 8, pp. 5859–5870, May 2023.
[242] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter, "Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification," *Frontiers Aging Neurosci.*, vol. 11, Jul. 2019, Art. no. 456892.
[243] J. Tobias Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014, *arXiv:1412.6806*.
[244] F. Eitel and K. Ritter, "Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification," in *Proc. 2nd Int. Workshop Interpretability Mach. Intell. Med. Image Comput. Multimodal Learn. Clin. Decision Support*, Shenzhen, China. Cham, Switzerland: Springer, Oct. 17, 2019, pp. 3–11.
[245] F. Dubost, P. Yilmaz, H. Adams, G. Bortsova, M. A. Ikram, W. Niessen, M. Vernooij, and M. de Bruijne, "Enlarged perivascular spaces in brain MRI: Automated quantification in four regions," *NeuroImage*, vol. 185, pp. 534–544, Jan. 2019.
[246] X. Wang, X. Liang, Z. Jiang, B. A. Nguchu, Y. Zhou, Y. Wang, H. Wang, Y. Li, Y. Zhu, F. Wu, J. Gao, and B. Qiu, "Decoding and mapping task states of the human brain via deep learning," *Human Brain Mapping*, vol. 41, no. 6, pp. 1505–1519, Apr. 2020.
[247] N. Gessert, S. Latus, Y. S. Abdelwahed, D. M. Leistner, M. Lutz, and A. Schlaefer, "Bioresorbable scaffold visualization in IVOCT images using CNNs and weakly supervised localization," *Proc. SPIE*, vol. 10949, pp. 606–612, Mar. 2019.
[248] K. Wickstrøm, M. Kampffmeyer, and R. Jenssen, "Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps," *Med. Image Anal.*, vol. 60, Feb. 2020, Art. no. 101619.
[249] A. Jamaludin, T. Kadir, and A. Zisserman, "SpineNet: Automated classification and evidence visualization in spinal MRIs," *Med. Image Anal.*, vol. 41, pp. 63–73, Oct. 2017.
[250] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.
[251] Z. Papanastasopoulos, R. K. Samala, H. Chan, L. M. Hadjiiski, C. Paramagul, M. A. Helvie, and C. H. Neal, "Explainable AI for medical imaging: Deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI," *Proc. SPIE*, vol. 11314, pp. 228–235, Mar. 2020.
[252] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3449–3457.
[253] A. Farrag, G. Gad, Z. M. Fadlullah, M. M. Fouda, and M. Alsabaan, "An explainable AI system for medical image segmentation with preserved local resolution: Mammogram tumor segmentation," *IEEE Access*, vol. 11, pp. 125543–125561, 2023.
[254] M. Chen, M. Zhang, L. Yin, L. Ma, R. Ding, T. Zheng, Q. Yue, S. Lui, and H. Sun, "Medical image foundation models in assisting diagnosis of brain tumors: A pilot study," *Eur. Radiol.*, vol. 34, no. 10, pp. 6667–6679, Apr. 2024.
[255] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2017, pp. 3319–3328.
[256] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," 2017, *arXiv:1711.06104*.
[257] R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu, S. Xu, S. Barb, A. Joseph, M. Shumski, J. Smith, A. B. Sood, G. S. Corrado, L. Peng, and D. R. Webster, "Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy," *Ophthalmology*, vol. 126, no. 4, pp. 552–564, Apr. 2019.
[258] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2017, pp. 3145–3153.
[259] K. H. Patil and M. N. Bhat, "Deep lift incorporated deep learning framework for classification of novel coronavirus (COVID-19) using computed tomography scan images," *Frontiers Health Informat.*, vol. 2024, pp. 6282–6299, Jun. 2024.
[260] W. Jin, X. Li, M. Fatehi, and G. Hamarneh, "Generating post-hoc explanation from deep neural networks for multi-modal medical image analysis tasks," *MethodsX*, vol. 10, Jul. 2023, Art. no. 102009.
[261] M. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.
[262] Md. S. Kamal, A. Northcote, L. Chowdhury, N. Dey, R. G. Crespo, and E. Herrera-Viedma, "Alzheimer's patient analysis using image and gene expression data and explainable-AI to present associated genes," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–7, 2021.
[263] R. Ghnemat, S. Alodibat, and Q. Abu Al-Haija, "Explainable artificial intelligence (XAI) for deep learning based medical imaging classification," *J. Imag.*, vol. 9, no. 9, p. 177, Aug. 2023.
[264] M. M. Ahsan, R. Nazim, Z. Siddique, and P. Huebner, "Detection of COVID-19 patients from CT scan and chest X-ray data using modified MobileNetV2 and LIME," *Healthcare*, vol. 9, no. 9, p. 1099, Aug. 2021.
[265] M. Rucco, G. Viticchi, and L. Falsetti, "Towards personalized diagnosis of glioblastoma in fluid-attenuated inversion recovery (FLAIR) by topological interpretable machine learning," *Mathematics*, vol. 8, no. 5, p. 770, May 2020.
[266] S. Alkhalaf, F. Alturise, A. A. Bahaddad, B. M. E. Elnaim, S. Shabana, S. Abdel-Khalek, and R. F. Mansour, "Adaptive Aquila optimizer with explainable artificial intelligence-enabled cancer diagnosis on medical imaging," *Cancers*, vol. 15, no. 5, p. 1492, Feb. 2023.
[267] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017, *arXiv:1705.07874*.
[268] D. Fudenberg, *Game Theory*. Cambridge, MA, USA: MIT Press, 1991.

[269] J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, 1947.
[270] B. H. M. van der Velden, M. H. A. Janse, M. A. A. Ragusi, C. E. Loo, and K. G. A. Gilhuijs, "Volumetric breast density estimation on MRI using explainable deep learning regression," *Sci. Rep.*, vol. 10, no. 1, p. 18095, Oct. 2020.
[271] N. Q. K. Le, Q. H. Kha, V. H. Nguyen, Y.-C. Chen, S.-J. Cheng, and C.-Y. Chen, "Machine learning-based radiomics signatures for EGFR and Kras mutations prediction in non-small-cell lung cancer," *Int. J. Mol. Sci.*, vol. 22, no. 17, p. 9254, Aug. 2021.
[272] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, Apr. 2018, pp. 1–11.
[273] S. H. P. Abeyagunasekera, Y. Perera, K. Chamara, U. Kaushalya, P. Sumathipala, and O. Senaweera, "LISA: Enhance the explainability of medical images unifying current XAI techniques," in *Proc. IEEE 7th Int. Conf. Converg. Technol. (I2CT)*, Apr. 2022, pp. 1–9.
[274] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harv. JL Tech.*, vol. 31, p. 841, Oct. 2017.
[275] D. Lenis, D. Major, M. Wimmer, A. Berg, G. Sluiter, and K. Bühler, "Domain aware medical image classifier interpretation by counterfactual impact analysis," in *Proc. 23rd Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, Lima, Peru. Cham, Switzerland: Springer, Jul. 2020, pp. 315–325.
[276] B. Kim, M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. Viégas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2017, pp. 2668–2677.
[277] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.
[278] A. Janik, J. D. Dodd, G. Ifrim, K. Sankaran, and K. M. Curran, "Interpretability of a deep learning model in the application of cardiac MRI segmentation with an ACDC challenge dataset," *Proc. SPIE*, vol. 11596, pp. 861–872, Feb. 2021.
[279] J. R. Clough, I. Öksüz, E. Puyol-Antón, B. Ruijsink, A. P. King, and J. A. Schnabel, "Global and local interpretability for cardiac MRI classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Jan. 2019, pp. 656–664.
[280] P. Gamble, R. Jaroensri, H. Wang, F. Tan, M. Moran, T. Brown, I. Flament-Auvigne, E. A. Rakha, M. Toss, D. J. Dabbs, P. Regitnig, N. Olson, J. H. Wren, C. Robinson, G. S. Corrado, L. H. Peng, Y. Liu, C. H. Mermel, D. F. Steiner, and P.-H.-C. Chen, "Determining breast cancer biomarker status and associated morphological features using deep learning," *Commun. Med.*, vol. 1, no. 1, p. 14, Jul. 2021.
[281] A. Lucieri, M. N. Bajwa, S. Alexander Braun, M. I. Malik, A. Dengel, and S. Ahmed, "On interpretability of deep learning based skin lesion classifiers using concept activation vectors," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–10.
[282] O. Li, H. Liu, C. Chen, and C. Rudin, "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, Jan. 2017, pp. 1–16.
[283] C. Chen, O. Li, C. Tao, A. J. Barnett, J. K. Su, and C. Rudin, "This looks like that: Deep learning for interpretable image recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Jan. 2018, pp. 1–11.
[284] J. Donnelly, A. J. Barnett, and C. Chen, "Deformable ProtoPNet: An interpretable image classifier using deformable prototypes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10255–10265.
[285] P. Rath-Manakidis, F. Strothmann, T. Glasmachers, and L. Wiskott, "ProtoP-OD: Explainable object detection with prototypical parts," 2024, *arXiv:2402.19142*.
[286] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo, "Benchmarking and survey of explanation methods for black box models," *Data Mining Knowl. Discovery*, vol. 37, no. 5, pp. 1719–1778, Sep. 2023.
[287] R. Correa, J. Jason Jeong, B. Patel, H. Trivedi, J. W. Gichoya, and I. Banerjee, "Two-step adversarial debiasing with partial learning - medical image case-studies," 2021, *arXiv:2111.08711*.
[288] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[289] D. Fan, Y. Wu, and X. Li, "On the fairness of swarm learning in skin lesion classification," in *Proc. Clin. Image-Based Procedures, Distrib. Collaborative Learn., Artif. Intell. Combating COVID-19 Secure Privacy-Preserving Mach. Learn.*, Strasbourg, France. Cham, Switzerland: Springer, Jan. 2021, pp. 120–129.
[290] J.-F. Rajotte, S. Mukherjee, C. Robinson, A. Ortiz, C. West, J. M. L. Ferres, and R. T. Ng, "Reducing bias and increasing utility by federated generative modeling of medical images using a centralized adversary," in *Proc. Conf. Inf. Technol. Social Good*, Sep. 2021, pp. 79–84.
[291] C. Wachinger and M. Reuter, "Domain adaptation for Alzheimer's disease diagnostics," *NeuroImage*, vol. 139, pp. 470–479, Oct. 2016.
[292] C. Xue, Q. Dou, X. Shi, H. Chen, and P.-A. Heng, "Robust learning at noisy labeled medical images: Applied to skin lesion classification," in *Proc. IEEE 16th Int. Symp. Biomed. Imag.(ISBI)*, Apr. 2019, pp. 1280–1283.
[293] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl. Inf. Syst.*, vol. 33, no. 1, pp. 1–33, Oct. 2012.
[294] M. M. Afzal, M. O. Khan, and S. Mirza, "Towards equitable kidney tumor segmentation: Bias evaluation and mitigation," in *Proc. Mach. Learn. Health (ML4H)*, 2023, pp. 13–26.
[295] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.
[296] Y. Zong, Y. Yang, and T. Hospedales, "MEDFAIR: Benchmarking fairness for medical imaging," 2022, *arXiv:2210.01725*.
[297] I. Ktena, O. Wiles, I. Albuquerque, S.-A. Rebuffi, R. Tanno, A. G. Roy, S. Azizi, D. Belgrave, P. Kohli, T. Cemgil, A. Karthikesalingam, and S. Gowal, "Generative models improve fairness of medical classifiers under distribution shifts," *Nature Med.*, vol. 30, no. 4, pp. 1166–1173, Apr. 2024.
[298] P. Burlina, N. Joshi, W. Paul, K. D. Pacheco, and N. M. Bressler, "Addressing artificial intelligence bias in retinal diagnostics," *Transl. Vis. Sci. Technol.*, vol. 10, no. 2, p. 13, Feb. 2021.
[299] G. Blanchard, G. Lee, and C. Scott, "Generalizing from several related classification tasks to a new unlabeled sample," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, Dec. 2011, pp. 2178–2186.
[300] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020.
[301] Q. Zhao, E. Adeli, and K. M. Pohl, "Training confounder-free deep learning models for medical applications," *Nature Commun.*, vol. 11, no. 1, p. 6010, Nov. 2020.
[302] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proc. 3rd Innov. Theor. Comput. Sci. Conf.*, New York, NY, USA, Jan. 2012, pp. 214–226.
[303] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness constraints: A flexible approach for fair classification," *J. Mach. Learn. Res.*, vol. 20, no. 75, pp. 1–42, Jan. 2019.
[304] C. Yang, Y. Sheng, P. Dong, Z. Kong, Y. Li, P. Yu, L. Yang, X. Lin, and Y. Wang, "Fast and fair medical AI on the edge through neural architecture search for hybrid vision models," in *Proc. IEEE/ACM Int. Conf. Comput. Aided Design (ICCAD)*, Oct. 2023, pp. 1–9.
[305] C.-H. Chiu, H. Chung, Y.-J. Chen, Y. Shi, and T.-Y. Ho, "Toward fairness through fair multi-exit framework for dermatological disease diagnosis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Jan. 2023, pp. 97–107.
[306] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2017, pp. 2564–2572.
[307] S. Feuerriegel, M. Dolata, and G. Schwabe, "Fair AI: Challenges and opportunities," *Bus. Inf. Syst. Eng.*, vol. 62, pp. 379–384, Aug. 2020.
[308] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, Jan. 2016, pp. 1–16.
[309] P. K. Lohia, K. Natesan Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, and R. Puri, "Bias mitigation post-processing for individual and group fairness," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2847–2851.
[310] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Sep. 2017, pp. 5680–5689.

[311] M. Jakesch, Z. Buçinca, S. Amershi, and A. Olteanu, "How different groups prioritize ethical values for responsible AI," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, Jun. 2022, pp. 310–323.
[312] A. Buhmann and C. Fieseler, "Towards a deliberative framework for responsible innovation in artificial intelligence," *Technol. Soc.*, vol. 64, Feb. 2021, Art. no. 101475.
[313] V. Dignum, "Responsible artificial intelligence: Designing AI for human values," *Int. Telecommun. Union*, vol. 2017, no. 1, pp. 1–8, 2017.
[314] U. Sivarajah, Y. Wang, H. Olya, and S. Mathew, "Responsible artificial intelligence (AI) for digital health and medical analytics," *Inf. Syst. Frontiers*, vol. 25, no. 6, pp. 2117–2122, Dec. 2023.
[315] P. Kumar, Y. K. Dwivedi, and A. Anand, "Responsible artificial intelligence (AI) for value formation and market performance in healthcare: The mediating role of patient's cognitive engagement," *Inf. Syst. Frontiers*, vol. 25, no. 6, pp. 2197–2220, Dec. 2023.
[316] F. Quazi, "Ethics & responsible AI in healthcare," *SSRN Electron. J.*, 2024.
[317] D. R. M. Lukkien, H. H. Nap, H. P. Buimer, A. Peine, W. P. C. Boon, J. C. F. Ket, M. M. N. Minkman, and E. H. M. Moors, "Toward responsible artificial intelligence in long-term care: A scoping review on practical approaches," *Gerontologist*, vol. 63, no. 1, pp. 155–168, Jan. 2023.
[318] M. I. Merhi, "An assessment of the barriers impacting responsible artificial intelligence," *Inf. Syst. Frontiers*, vol. 25, no. 3, pp. 1147–1160, Jun. 2023.
[319] Q. Lu, L. Zhu, J. Whittle, and X. Xu, *Responsible AI: Best Practices for Creating Trustworthy AI Systems*. Reading, MA, USA: Addison-Wesley, 2023.
[320] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.
[321] S. M. Muddamsetty, M. N. S. Jahromi, and T. B. Moeslund, "Expert level evaluations for explainable AI (XAI) methods in the medical domain," in *Proc. Int. Conf. Pattern Recognit.* Cham, Switzerland: Springer, Jan. 2021, pp. 35–46.
[322] L. Coroama and A. Groza, "Evaluation metrics in explainable artificial intelligence (XAI)," in *Proc. Int. Conf. Adv. Res. Technol., Inf., Innov. Sustainability*. Cham, Switzerland: Springer, Jan. 2022, pp. 401–413.
[323] M. Mozolewski, S. Bobek, and G. J. Nalepa, "Visual explanations and perturbation-based fidelity metrics for feature-based models," in *Proc. Int. Conf. Comput. Sci.*, Jan. 2024, pp. 294–309.
[324] J. M. Darias, B. Bayrak, M. Caro-Martínez, B. Díaz-Agudo, and J. A. Recio-García, "An empirical analysis of user preferences regarding XAI metrics," in *Proc. Int. Conf. Case-Based Reasoning*. Cham, Switzerland: Springer, Jan. 2024, pp. 96–110.
[325] C. Agarwal, E. Saxena, S. Krishna, M. Pawelczyk, N. Johnson, I. Puri, M. Žitnik, and H. Lakkaraju, "OpenXAI: Towards a transparent evaluation of model explanations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, Jan. 2022, pp. 15784–15799.
[326] N. Fouladgar, M. Alirezaie, and K. Framling, "Metrics and evaluations of time series explanations: An application in affect computing," *IEEE Access*, vol. 10, pp. 23995–24009, 2022.
[327] P. Banerjee and R. P. Barnwal, "Methods and metrics for explaining artificial intelligence models: A review," in *Explainable AI: Foundations, Methodologies and Applications*. Cham, Switzerland: Springer, 2023, pp. 61–88.
[328] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *Int. J. Comput. Vis.*, vol. 126, no. 10, pp. 1084–1102, Oct. 2018.
[329] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, and S. Lapuschkin, "Towards best practice in explaining neural network decisions with LRP," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–7.
[330] I. Rio-Torto, T. Gonçalves, J. S. Cardoso, and L. F. Teixeira, "On the suitability of B-cos networks for the medical domain," in *Proc. IEEE Int. Symp. Biomed. Imag. (ISBI)*, May 2024, pp. 1–5.
[331] G. Papanastasiou, N. Dikaios, J. Huang, C. Wang, and G. Yang, "Is attention all you need in medical image analysis? A review," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 3, pp. 1398–1411, Mar. 2024.
[332] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jan. 2017, pp. 3681–3688.
[333] J. Theiner, E. Müller-Budack, and R. Ewerth, "Interpretable semantic photo geolocation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1474–1484.
[334] L. Arras, A. Osman, and W. Samek, "CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations," *Inf. Fusion*, vol. 81, pp. 14–40, May 2022.
[335] M. Springenberg, A. Frommholz, M. Wenzel, E. Weicken, J. Ma, and N. Strodthoff, "From modern CNNs to vision transformers: Assessing the performance, robustness, and classification strategies of deep learning models in histopathology," *Med. Image Anal.*, vol. 87, Jul. 2023, Art. no. 102809.
[336] V. Srinivasan, N. Strodthoff, J. Ma, A. Binder, K.-R. Müller, and W. Samek, "To pretrain or not? A systematic analysis of the benefits of pretraining in diabetic retinopathy," *PLoS ONE*, vol. 17, no. 10, Oct. 2022, Art. no. e0274291.
[337] K. K. Wickstrøm, E. A. Østmo, K. Radiya, K. Ø. Mikalsen, M. C. Kampffmeyer, and R. Jenssen, "A clinically motivated self-supervised approach for content-based image retrieval of CT liver images," *Computerized Med. Imag. Graph.*, vol. 107, Jul. 2023, Art. no. 102239.
[338] U. Bhatt, A. Weller, and J. M. F. Moura, "Evaluating and aggregating feature-based model explanations," 2020, *arXiv:2005.00631*.
[339] P. Komorowski, H. Baniecki, and P. Biecek, "Towards evaluating explanations of vision transformers for medical imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 3726–3732.
[340] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2660–2673, Nov. 2017.
[341] B. Hu, B. Vasu, and A. Hoogs, "X-MIR: EXplainable medical image retrieval," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1544–1554.
[342] I. E. Nielsen, R. P. Ramachandran, N. Bouaynaya, H. M. Fathallah-Shaykh, and G. Rasool, "EvalAttAI: A holistic approach to evaluating attribution maps in robust and non-robust models," *IEEE Access*, vol. 11, pp. 82556–82569, 2023.
[343] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Jan. 2018, pp. 1–14.
[344] A. Sadafi, O. Adonkina, A. Khakzar, P. Lienemann, R. M. Hehr, D. Rueckert, N. Navab, and C. Marr, "Pixel-level explanation of multiple instance learning models in biomedical single cell images," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Cham, Switzerland: Springer, Jan. 2023, pp. 170–182.
[345] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci, "A consistent and efficient evaluation strategy for attribution methods," 2022, *arXiv:2202.00449*.
[346] V. Lamprou, A. Kallipolitis, and I. Maglogiannis, "On the evaluation of deep learning interpretability methods for medical images under the scope of faithfulness," *Comput. Methods Programs Biomed.*, vol. 253, Aug. 2024, Art. no. 108238.
[347] A. Kallipolitis, P. Yfantis, and I. Maglogiannis, "Improving explainability results of convolutional neural networks in microscopy images," *Neural Comput. Appl.*, vol. 35, no. 29, pp. 21535–21553, Oct. 2023.
[348] M. Gallo, V. Krajňanský, R. Nenutil, P. Holub, and T. Brázdil, "Shedding light on the black box of a neural network used to detect prostate cancer in whole slide images by occlusion-based explainability," *New Biotechnol.*, vol. 78, pp. 52–67, Dec. 2023.
[349] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," 2018, *arXiv:1806.08049*.
[350] E. Doumard, J. Aligon, E. Escriva, J.-B. Excoffier, P. Monsarrat, and C. Soulé-Dupuy, "A quantitative approach for the comparison of additive local explanation methods," *Inf. Syst.*, vol. 114, Mar. 2023, Art. no. 102162.
[351] J. Sun, W. Shi, F. O. Giuste, Y. S. Vaghani, L. Tang, and M. D. Wang, "Improving explainable AI with patch perturbation-based evaluation pipeline: A COVID-19 X-ray image analysis case study," *Sci. Rep.*, vol. 13, no. 1, p. 19488, Nov. 2023.
[352] A. Binder, L. Weber, S. Lapuschkin, G. Montavon, K.-R. Müller, and W. Samek, "Shortcomings of top-down randomization-based sanity checks for evaluations of deep neural network explanations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 16143–16152.
[353] M. A. Kadir, A. Mohamed Selim, M. Barz, and D. Sonntag, "A user interface for explaining machine learning model explanations," in *Proc. 28th Int. Conf. Intell. User Interfaces*, Mar. 2023, pp. 59–63.

[354] W.-J. Nam, S. Gur, J. Choi, L. Wolf, and S.-W. Lee, "Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Apr. 2020, pp. 2501–2508.

[355] N. Gnanavel, P. Inparaj, N. Sritharan, D. Meedeniya, and P. Yogarajah, "Interpretable cervical cell classification: A comparative analysis," in *Proc. 4th Int. Conf. Adv. Res. Comput. (ICARC)*, Feb. 2024, pp. 7–12.

[356] V. Pitroda, M. M. Fouda, and Z. M. Fadlullah, "An explainable AI model for interpretable lung disease classification," in *Proc. IEEE Int. Conf. Internet Things Intell. Syst. (IoTaIS)*, Nov. 2021, pp. 98–103.

[357] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, Oct. 2018, pp. 9505–9515.

[358] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini, "Factual and counterfactual explanations for black box decision making," *IEEE Intell. Syst.*, vol. 34, no. 6, pp. 14–23, Nov. 2019.

[359] M. Velmurugan, C. Ouyang, C. Moreira, and R. Sindhgatta, "Developing a fidelity evaluation approach for interpretable machine learning," 2021, *arXiv:2106.08492*.

[360] A. Buliga, M. Vazifehdoostirani, L. Genga, X. Lu, R. Dijkman, C. D. Francescomarino, C. Ghidini, and H. A. Reijers, "Uncovering patterns for local explanations in outcome-based predictive process monitoring," in *Proc. Int. Conf. Bus. Process Manage.* Cham, Switzerland: Springer, Jan. 2024, pp. 363–380.

[361] T.-H. Huang, A. Metzger, and K. Pohl, "Counterfactual explanations for predictive business process monitoring," in *Proc. Eur., Medit., Middle Eastern Conf. Inf. Syst.* Cham, Switzerland: Springer, Jan. 2022, pp. 399–413.

[362] D. Alvarez-Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, Dec. 2018, pp. 7786–7795.

[363] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 607–617.

[364] S. Verma, V. Boonsanong, M. Hoang, K. E. Hines, J. P. Dickerson, and C. Shah, "Counterfactual explanations and algorithmic recourses for machine learning: A review," 2020, *arXiv:2010.10596*.

[365] S. Verma, V. Boonsanong, M. Hoang, K. Hines, J. Dickerson, and C. Shah, "Counterfactual explanations and algorithmic recourses for machine learning: A review," *ACM Comput. Surveys*, vol. 56, no. 12, pp. 1–42, Dec. 2024.

[366] S. Singla, M. Eslami, B. Pollack, S. Wallace, and K. Batmanghelich, "Explaining the black-box smoothly—A counterfactual approach," *Med. Image Anal.*, vol. 84, Feb. 2023, Art. no. 102721.

[367] A. Lamiable, T. Champetier, F. Leonardi, E. Cohen, P. Sommer, D. Hardy, N. Argy, A. Massougbodji, E. Del Nery, G. Cottrell, Y.-J. Kwon, and A. Genovesio, "Revealing invisible cell phenotypes with conditional generative modeling," *Nature Commun.*, vol. 14, no. 1, p. 6386, Oct. 2023.

[368] L. R. Koetzier, J. Wu, D. Mastrodicasa, A. Lutz, M. Chung, W. A. Koszek, J. Pratap, A. S. Chaudhari, P. Rajpurkar, M. P. Lungren, and M. J. Willemink, "Generating synthetic data for medical imaging," *Radiology*, vol. 312, no. 3, Sep. 2024, Art. no. e232471.

[369] Y. Xu, L. Sun, W. Peng, S. Jia, K. Morrison, A. Perer, A. Zandifar, S. Visweswaran, M. Eslami, and K. Batmanghelich, "MedSyn: Text-guided anatomy-aware synthesis of high-fidelity 3-D CT images," *IEEE Trans. Med. Imag.*, vol. 43, no. 10, pp. 3648–3660, Oct. 2024.

[370] A. Ghandeharioun, B. Kim, C.-L. Li, B. Jou, B. Eoff, and R. W. Picard, "DISSECT: Disentangled simultaneous explanations via concept traversals," 2021, *arXiv:2105.15164*.

[371] C. Patrício, J. C. Neves, and L. F. Teixeira, "Explainable deep learning methods in medical image classification: A survey," 2022, *arXiv:2205.04766*.

[372] A. V. Looveren and J. Klaise, "Interpretable counterfactual explanations guided by prototypes," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, Jan. 2021, pp. 650–665.

[373] D. Mahajan, C. Tan, and A. Sharma, "Preserving causal constraints in counterfactual explanations for machine learning classifiers," 2019, *arXiv:1912.03277*.

[374] W. Taylor-Melanson, Z. Sadeghi, and S. Matwin, "Causal generative explainers using counterfactual inference: A case study on the morpho-MNIST dataset," *Pattern Anal. Appl.*, vol. 27, no. 3, p. 89, Sep. 2024.

[375] J. L. Soto, E. Z. Uriguen, and X. De Carlos Garcia, "Real-time, model-agnostic and user-driven counterfactual explanations using autoencoders," *Appl. Sci.*, vol. 13, no. 5, p. 2912, Feb. 2023.

[376] S. Khorram and L. Fuxin, "Cycle-consistent counterfactuals by latent transformations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10193–10202.

[377] A. Ghorbani, J. Wexler, J. Zou, and B. Kim, "Towards automatic concept-based explanations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Jan. 2019, pp. 1–7.

[378] C. Yeh, B. Kim, S. Ö. Arık, C. Li, T. Pfister, and P. Ravikumar, "On completeness-aware concept-based explanations in deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Jan. 2019, pp. 20554–20565.

[379] C. Ma, B. Zhao, C. Chen, and C. Rudin, "This looks like those: Illuminating prototypical concepts using multiple visualizations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, Jan. 2023, pp. 1–8.

[380] F. Foscarin, K. Hoedt, V. Praher, A. Flexer, and G. Widmer, "Concept-based techniques for 'musicologist-friendly' explanations in a deep music classifier," 2022, *arXiv:2208.12485*.

[381] M. Kowal, R. P. Wildes, and K. G. Derpanis, "Visual concept connectome (VCC): Open world concept discovery and their interlayer connections in deep models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 10895–10905.

[382] T. Chanda et al., "Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma," *Nature Commun.*, vol. 15, no. 1, p. 524, Jan. 2024.

[383] M. Du, F. Yang, N. Zou, and X. Hu, "Fairness in deep learning: A computational perspective," *IEEE Intell. Syst.*, vol. 36, no. 4, pp. 25–34, Jul. 2021.

[384] P. Awasthi, M. Kleindeßner, and J. Morgenstern, "Equalized odds postprocessing under imperfect group information," in *Proc. Int. Conf. Artif. Intell. Statist.*, Jun. 2020, pp. 1770–1780.

[385] N. Goel, M. Yaghini, and B. Faltings, "Non-discriminatory machine learning through convex fairness criteria," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Dec. 2018, p. 116.

[386] S. Jung, D. Lee, T. Park, and T. Moon, "Fair feature distillation for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12110–12119.

[387] D. Saha, C. Schumann, D. C. McElfresh, J. P. Dickerson, M. L. Mazurek, and M. C. Tschantz, "Measuring non-expert comprehension of machine learning fairness metrics," in *Proc. 37th Int. Conf. Mach. Learn.*, vol. 119, H. Daumé and A. Singh, Eds., Jul. 2020, pp. 8377–8387.

[388] H. Narasimhan, A. Cotter, M. R. Gupta, and S. Wang, "Pairwise fairness for ranking and regression," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Apr. 2020, pp. 5248–5255.

[389] A. Beutel, J. Chen, T. Doshi, H. Qian, A. Woodruff, C. Luu, P. Kreitmann, J. Bischof, and E. H. Chi, "Putting fairness principles into practice: Challenges, metrics, and improvements," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, New York, NY, USA, Jan. 2019, pp. 453–459.

[390] J. Huang, G. Galal, M. Etemadi, and M. Vaidyanathan, "Evaluation and mitigation of racial bias in clinical machine learning models: Scoping review," *JMIR Med. Informat.*, vol. 10, no. 5, May 2022, Art. no. e36388.

[391] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, and A. C. Cosentini, "A clarification of the nuances in the fairness metrics landscape," *Sci. Rep.*, vol. 12, no. 1, p. 4209, Mar. 2022.

[392] L. Cohausz, J. Kappenberger, and H. Stuckenschmidt, "What fairness metrics can really tell you: A case study in the educational domain," in *Proc. 14th Learn. Anal. Knowl. Conf.*, New York, NY, USA, Mar. 2024, pp. 792–799.

[393] A. Ghosh, A. Shanbhag, and C. Wilson, "FairCanary: Rapid continuous explainable fairness," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, New York, NY, USA, Jul. 2022, pp. 307–316.

[394] S. Du, B. Hers, N. Bayasi, G. Hamarneh, and R. Garbi, "FairDisCo: Fairer AI in dermatology via disentanglement contrastive learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Jan. 2023, pp. 185–202.

[395] M. Wang and W. Deng, "Mitigating bias in face recognition using skewness-aware reinforcement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9319–9328.

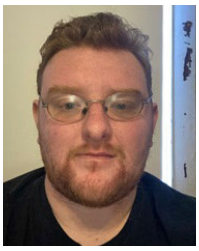
[396] T. Ohki, Y. Sato, M. Nishigaki, and K. Ito, "LabellessFace: Fair metric learning for face recognition without attribute labels," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2024, pp. 1–8.

[397] A. Atzori, P. Cosseddu, G. Fenu, and M. Marras, "The impact of balancing real and synthetic data on accuracy and fairness in face recognition," 2024, *arXiv:2409.02867*.

[398] U. Ruby and V. Yendapalli, "Binary cross entropy with deep learning technique for image classification," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 4, pp. 5393–5397, Aug. 2020.
[399] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2017, pp. 1321–1330.
[400] J. Nixon, M. Dusenberry, G. Jerfel, T. Nguyen, J. Z. Liu, L. Zhang, and D. Tran, "Measuring calibration in deep learning," in *Proc. CVPR Workshops*, vol. 2, Jan. 2019, pp. 1–4.
[401] A. Brahmbhatt, V. Rathore, and P. Singla, "Towards fair and calibrated models," 2023, *arXiv:2310.10399*.
[402] M. Lin, T. Li, Y. Yang, G. Holste, Y. Ding, S. H. Van Tassel, K. Kovacs, G. Shih, Z. Wang, Z. Lu, F. Wang, and Y. Peng, "Improving model fairness in image-based computer-aided diagnosis," *Nature Commun.*, vol. 14, no. 1, p. 6261, Oct. 2023.
[403] C. Hazirbas, J. Bitton, B. Dolhansky, J. Pan, A. Gordo, and C. C. Ferrer, "Towards measuring fairness in AI: The casual conversations dataset," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 4, no. 3, pp. 324–332, Jul. 2022.
[404] A. Ferrara, F. Bonchi, F. Fabbri, F. Karimi, and C. Wagner, "Bias-aware ranking from pairwise comparisons," *Data Mining Knowl. Discovery*, vol. 38, no. 4, pp. 2062–2086, Jul. 2024.
[405] Z. Dodevska, S. Radovanović, A. Petrović, and B. Delibašić, "When fairness meets consistency in AHP pairwise comparisons," *Mathematics*, vol. 11, no. 3, p. 604, Jan. 2023.
[406] Y. Tian, M. Shi, Y. Luo, A. Kouhana, T. Elze, and M. Wang, "FairSeg: A large-scale medical image segmentation dataset for fairness learning using segment anything model with fair error-bound scaling," 2023, *arXiv:2311.02189*.
[407] M. Masroor, T. Hassan, Y. Tian, K. Wells, D. Rosewarne, T.-T. Do, and G. Carneiro, "Fair distillation: Teaching fairness from biased teachers in medical imaging," 2024, *arXiv:2411.11939*.
[408] N. Littlefield, S. Amirian, J. Biehl, E. G. Andrews, M. Kann, N. Myers, L. Reid, A. J. Yates, B. J. McGrory, B. Parmanto, T. M. Seyler, J. F. Plate, H. H. Rashidi, and A. P. Tafti, "Generative AI in orthopedics: An explainable deep few-shot image augmentation pipeline for plain knee radiographs and Kellgren–Lawrence grading," *J. Amer. Med. Inform. Assoc.*, vol. 31, no. 11, pp. 2668–2678, Nov. 2024.
[409] Osteoarthritis Initiative. (2024). *Osteoarthritis Initiative (OAI) Dataset*. Accessed: Sep. 1, 2024. [Online]. Available: https://nda.nih.gov/oai
[410] N. Littlefield, J. F. Plate, K. R. Weiss, I. Lohse, A. Chhabra, I. A. Siddiqui, Z. Menezes, G. Mastorakos, S. Mehul Thakar, M. Abedian, M. F. Gong, L. A. Carlson, H. Moradi, S. Amirian, and A. P. Tafti, "Learning unbiased image segmentation: A case study with plain knee radiographs," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Oct. 2023, pp. 1–5.

SOHEYLA AMIRIAN received the Ph.D. degree. She was a Faculty Fellow at the Institute for Artificial Intelligence and a Faculty Lecturer at the School of Computing, University of Georgia, for three years. She is currently an Assistant Professor at the Seidenberg School of Computer Science and Information Systems, Pace University. She leads the Applied Machine Intelligence Initiatives and Education (AMIIE) Laboratory, collaborating with a multidisciplinary team of faculty, students, and investigators to design, build, validate, and deploy AI algorithms in various real-world applications, including public health, imaging informatics, and AI-powered education. She authored over 25 peer-reviewed publications. She has served as a Program Committee Member at IEEE ICHI and the Co-Chair of Research Tracks at the World Congress in Computer Science, Computer Engineering, and Applied Computing (CSCE) and CSCI. She has received numerous accolades, including the 2019 International Conference on Computational Science and Computational Intelligence (CSCI) Outstanding Achievement Award, the 2021 UGA Outstanding Teaching Assistant Award, the NVIDIA GPU Award, and the ACM Richard TAPIA Conference Scholarship, in 2020 and 2022. In addition, she was a Finalist of the 2020 NCWIT Collegiate Award. She has organized several conferences and tutorials on computational intelligence, such as ISVC and IEEE ICHI. In 2023, she received the IEEE Atlanta Section Outstanding Educator Award. Most recently, she was awarded a grant from the NIH/National Institute on Aging (NIA) and a Faculty Fellowship from the Helene T. and Grant M. Wilson Center.

FENGYI GAO received the M.Sc. degree in health informatics from the University of Pittsburgh. She is currently a Research Assistant with the HexAI Research Laboratory, University of Pittsburgh. Her primary research interests include advanced artificial intelligence (AI) to tackle real-world medical and scientific challenges. The bulk of her research has been focused on developing and validating data mining and machine learning models using structured and unstructured clinical data to address clinical problems and improve patient care.



NICKOLAS LITTLEFIELD (Graduate Student Member, IEEE) received the M.S. degree. He is currently pursuing the Ph.D. degree with the Intelligent Systems Program (ISP), University of Pittsburgh. He is also an Active Researcher at the Pitt Health and Explainable AI Research Laboratory (Pitt HexAI) and the Center for Computational Pathology and AI Excellence (CPACE). He has already authored several scientific papers and abstracts in top-notch conferences and journals, including *Scientific Reports* (Nature) and *Journal of the American Medical Informatics Association*, and has organized a list of tutorials and AI summer schools for education for all. His research primarily targets enhancing the explainability and fairness of AI systems mainly in medical image analysis and building safe and responsible AI systems.

JONATHAN H. HILL received the D.P.S. degree in computing from Pace University, New York, NY, USA, and the M.B.A. degree in management from Baruch College, CUNY. He serves as the Interim Provost and the Executive Vice President of Academic Affairs at Pace University, overseeing academic operations across campuses in New York City, Pleasantville, and White Plains. With over 30 years of experience in academia and industry, he is known for empowering teams to deliver exceptional results and driving innovation in education and research. During his tenure as the Dean of the Seidenberg School of Computer Science and Information Systems, he led a revitalization of the school by introducing cutting-edge programs in artificial intelligence, data science, cybersecurity, and human-centered design, while fostering research excellence and securing over $10 million in faculty grant funding. He has a strong track record in cultivating international collaborations, corporate partnerships, and student-centered initiatives, such as the Seidenberg Scholars Program.

ADOLPH J. YATES JR. has been at the University of Pittsburgh for 20 years, where he serves as a Professor and the Vice Chair for Quality within the Department of Orthopaedic Surgery. He is also the Chief of Orthopaedics at UPMC Shadyside specializing in adult reconstruction. His undergraduate concentration at Harvard was in economics. From there he matriculated to Johns Hopkins where he spent ten years through medical school, residency, and then on the faculty. Before returning to Pittsburgh, he was on the faculty of OHSU, WVU, and the University of Rochester. His Laboratory research has focused on the restoration of cartilage. His clinical research is centered on EBM including co-authorship of multiple guidelines including ones for prevention of VTE and surgical treatment of knee OA. His publications have addressed value-based purchasing and performance measures and their impact on equitable access to care. He has served on the FDA device panel and MEDCAC. He served six years on the Surgical Committee for the NQF as well as numerous technical expert panels for CMS/Acumen/Yale-CORE helping to formulate performance and cost measures within MACRA and surgeon/hospital performance programs.

JOHANNES F. PLATE received the Ph.D. degree. He is currently an Associate Professor of Orthopaedic Surgery at the University of Pittsburgh School of Medicine. He is also the Director of Adult Reconstruction Research and the Clinical Associate Director at the University of Pittsburgh HexAI Research Laboratory. He is an Orthopaedic Surgeon, fellowship-trained in primary and complex total hip and knee replacements. He is certified by the American Board of Orthopaedic Surgery. He has published over 100 journal articles and book chapters and has presented research at national and international conferences. His research interests include value-based care, robotic techniques for hip and knee replacement, mitigating patient risk factors to improve surgical outcomes, and the treatment of prosthetic joint infections. He is a member of American Association of Hip and Knee Surgeons, American Academy of Orthopaedic Surgeons, and American Orthopaedic Association. He serves on the editorial board and is a reviewer for various orthopedic journals.

LIRON PANTANOWITZ received the M.D., Ph.D., and M.H.A. degrees. He is currently the Maud Menten Professor and the Chair of the Department of Pathology, UPMC, and the University of Pittsburgh. He is a pioneer in the field of pathology informatics. He has dedicated over two decades to working in this innovative field. He co-founded the *Journal of Pathology Informatics*, published several leading textbooks in this field, and helped establish the pathology informatics essential for residents (PIER) curriculum. His research interest includes advancing digital pathology and computational pathology. He is a member of the editorial board for *Journal of Applied Science and Computations*, *Cancer Cytopathology*, *Diagnostic Cytopathology*, *Acta Cytologica*, *Cytojournal*, and *Cytopathology*. He served as the President for the Association for Pathology Informatics (API) and the Digital Pathology Association (DPA). He has been listed several times as a top doctor in the Pittsburgh Magazine and Pittsburgh Business Times.

HOOMAN H. RASHIDI received the M.D. and M.S. degrees. He is currently the Associate Dean of AI in medicine and a Professor and the Endowed Chair of Lombardi-Shinozuka Experimental Pathology Research at the University of Pittsburgh School of Medicine. He is also the Executive Director of the Computational Pathology and AI Center of Excellence (CPACE). He combines his passion for patient care, research, and education with his unique training in bioinformatics and machine learning (ML) to create innovative new tools (i.e., MILO, STNG, and Pitt-GPT) and resources (Hematology Outlines and Cleveland Clinic's AI Course) that improve clinical practice, research, and education. He is also the Co-Founder and a Senior Editor of HematologyOutlines, a very popular digital hematology atlas that is used internationally and endorsed by American Society of Clinical Pathology.

AHMAD P. TAFTI received the Ph.D. degree. He is currently an Assistant Professor of health informatics with the Department of Health Information Management, School of Health and Rehabilitation Sciences, University of Pittsburgh, with secondary appointments at the University of Pittsburgh School of Medicine and the Intelligent Systems Program (ISP), School of Computing and Information. He is affiliated with the Center for AI Innovation in Medical Imaging (CAIIMI), and also serving as an Associate Member at the UPMC Hillman Cancer Center. He is leading all research and educational efforts at the Pitt HexAI Research Laboratory, conducting the health and explainable AI Podcast series, while he is also the Interim Director of Scientific Affairs within the Computational Pathology and AI Center of Excellence (CPACE). He has authored more than 60 peer-reviewed publications in AI-powered healthcare. He is a fellow of American Medical Informatics Association. Moreover, he is honored to serve their community as the Vice Chair of the IEEE Computer Society at Pittsburgh. He has a deep passion for AI-powered healthcare informatics and health data science with better patient diagnosis, prognosis, and treatment using large-scale multiple clinical data sources and advanced computational algorithms. He is the 2021 SiiM Imaging Informatics Innovator awardee and Oracle for Research (Eureka) awardee.