Advancing Explainable AI in Deep Learning for Medical Imaging: Enhancing Transparency, Trust, and Clinical Utility

M Bhavsingh[1]* and P. Hussain Basha[2]
{bhavsinghit@gmail.com[1], drhussain_p@pace.ac.in[2]}

Research Scholar, Department of Computer Science and Engineering, JNTU Kakinada, Andhra Pradesh, India[1]
Assistant professor, Department of Computer Science and Engineering, Pace institute of Technology and Science, Ongole, Andhra Pradesh, India[2]

Abstract. In particular, this review gives a complete study of Explainable Artificial Intelligence (XAI) in deep learning applied to medical imaging diagnostics, which are of key importance in AI powered clinical decisions. The study presents an exhaustive yet practical exploration of a range of XAI methodologies such as gradient based visualization techniques, perturbation-based models, attribution mechanisms, attention-based networks, surrogate models and hybrids in applying to understand and trust medical XAI. The explainability significantly increases model interpretability and clinician acceptance and its practical use case on disease detection, segmentation, and prognostic analytics are then discussed. In addition to main challenges such as balance accuracy with interpretability, data quality constraints, algorithmic bias and the regulatory barriers, the paper addresses how data holes can be addressed. Furthermore, areas in multimodal data fusion, human in the loop AI, privacy preserving learning and federated AI are also explored to see how they may increase the model robustness and scalability. Performance, interpretability, computational efficiency and clinical utility of XAI techniques are compared among each other, and are presented to aid in the selection of suitable models for specific medical imaging applications. Finally, the paper concludes with some future research directions also specifying, regulatory compliance, computational optimization, and clinician centered AI framework as the need for such trustworthy, interpretable AI solutions to be adopted panprecise medicine.

Keywords: Explainable AI, Deep Learning, Medical Imaging, Interpretability, Clinical Decision Support.

1 Introduction

Modern medicine significantly advanced through medical imaging because it provides opportunities to identify diseases early and track their advancement and supports treatment preparations. Medical practice depends heavily on four imaging modalities including X-rays together with computed tomography (CT) and magnetic resonance imaging (MRI) and ultrasound for gaining detailed pictures of both anatomical features and pathological conditions [1]. Medical imaging methods benefit from artificial intelligence (AI) deployment and deep learning technology specifically which helps doctors achieve better diagnosis quality and automatic task performance [2]. Remains challenging for deep learning models to provide transparent decision-making since their operational functions remain unclear to healthcare professionals [3]. To address these concerns, Explainable AI (XAI) emerges in the hope of offering mechanisms that facilitate the model interpretability, allowing healthcare professionals

ICITSM-Part I 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357772

to understand and validating AI driven diagnosis [4]. It is critical to ensure an accuracy balance and transparency achieved in medical imaging for the purpose of fostering trust in AI assisted clinical workflows and to inform decision making [5]. Despite significant advancements, several challenges impede the widespread adoption of interpretable AI in medical imaging. Among these challenges are the lack of transparency in such models, limited trust in clinicians as a consequence, as well as the issue of balancing between accuracy and interpretability [6]. Besides, data scarcity, quality of annotations and algorithmic biases create a substantial challenge in developing and validating the models [7]. Technically, more complexity arises from integration of data sources from imaging, genomics, and clinical records. In addition, contributing to the area of real time explainability in the critical care needs innovation [9]. The final factor is that such medical systems anchored on AI definitely need to conform to ethical and regulatory frameworks in order to keep AI driven medical systems fair, private and accountable [10].

1.1 Key Contributions of the Paper

This paper constitutes a Systematic review of XAI techniques for medical imaging diagnostics combined with deep learning, with the following key contributions.

- **Comprehensive Survey of XAI Techniques:** Considering existing explainability methodologies including the model-specific, the model-agnostic and the hybrid approaches to enhance the transparency of the deep learning models on medical imaging.
- **Evaluation of Practical Applications:** Describes disease detection, segmentation and prognosis use cases where explainable AI improves clinical decision making and model interpretability.
- **Analysis of Technical and Ethical Challenges:** Identifies key barriers such as data quality issues, algorithmic bias, and the trade-offs between accuracy and interpretability, offering a critical assessment of limitations in current explainability frameworks.
- **Exploration of Emerging Trends:** Investigates recent advancements in multimodal data integration, human-in-the-loop frameworks, and federated learning, highlighting promising directions for future research.
- **Comparative Analysis of Explainability Models:** A structured review compares current XAI methods based on their performance levels and their interpretability, computational efficiency, and clinical usefulness capabilities for assisting medical imaging task selection.
- **Recommendations for Future Research:** The article suggests methods to achieve ethical alignment while securing patient privacy through AI medical imaging while solving regulatory problems and trust issues for scalability.

The paper divides into two parts: Section 2 explains explainable AI principles and deep learning integration with medical imaging technology. Section 3 evaluates present day explainability methods across three main categories such as model-specific and model-agnostic and hybrid approaches. A review of disease detection practices along with segmentation along with predictive analytics applications takes place in Section 4 through clinical examples. The implementation of explainable AI models encounters barriers from technical limitations and adoption issues while also encountering problems from data-related sources according to

Section 5. The discussion of future directions in Section 6 explores three emerging trends about multimodal integration of systems and privacy-preserving AI methods as well as human-in-the-loop systems. A comparative study of current methods appears in Section 7 to evaluate both their performance capabilities as well as interpretability metrics between different methods. A summary of conclusions along with research direction suggestions for the future marks the termination of this paper in Section 8.

2 Foundations of Explainable AI Integrating with Deep Learning in Medical Imaging

2.1 Definition and Importance

The fundamental requirements for AI applications extend to deep learning in medical imaging due to need for both explainability and interpretability. The degree of human understanding in AI model decision processes is called interpretability but explainability describes the ability to show internal AI decision mechanisms [11]. Many clinicians view deep learning models as black boxes because their complicated nature creates challenges for prediction validation which then questions both safety and reliability in patient care [12]. Healthcare professionals hold back from implementing AI diagnostic tools because these systems lack interpretability which leads them to hesitate [13]. Medical imaging and healthcare facilities need explainable AI models because these features create trust between practitioners and patients while maintaining transparency in diagnostic operations [14].

2.2 Historical Context

AI applications in healthcare started with initial expert systems then transitioned into machine learning operating on healthcare datasets. During the 1960s AI served in medical diagnosis through expert systems which used predetermined rules that attempted to reproduce physician decision processes. The groundbreaking systems which emerged at that time showed limited clinical usefulness because they possessed neither learning capabilities nor adaptive features [15]. Machine learning entered the scene in the late twentieth century to allow AI models use statistical techniques on massive datasets thus boosting predictive accuracy [16]. The implementation of these models caused medical institutions to face two major hurdles because their decisions became harder to understand which made them hard to accept in standard clinical workflows [17]. The latest research has brought forward AI models which provide maintenance of both interpretability and predictive accuracy in their operations. Deep neural networks receive improved explainability through model-agnostic interpretability frameworks together with self-attentive deep learning models according to research conducted in [18]. These diagnostic methods enable medical staff to track the reasoning behind AI diagnoses therefore making AI more trustworthy and clinically relevant. XAI has gained prominence because it demonstrates the need for transparent AI applications in healthcare to create accountable medical decision tools [19].

2.3 Ethical and Legal Considerations

XAI for medical imaging demands closest evaluation of ethical and legal aspects that determine clinical deployments. AI-driven medical imaging systems require patient data privacy and security to be primary considerations because their operations depend on massive sensitive health-related information databases. The protection of personal data must follow standards based on GDPR and HIPAA directives to reduce exposure risks arising from unauthorized

access breaches [20]. Deep learning models present such complex systems that fault detection for diagnostic errors becomes difficult to trace during diagnostic processes. Medical imaging XAI implementation faces regulatory barriers because courts are actively debating which party bears responsibility when diagnostic issues occur [21]. This lack of legal precision forces institutions to determine between healthcare providers developers or deployment organizations [21]. Existing laws want improvement to provide proper regulations to AI technologies because modern standards do not fully handle complex artificial intelligence operational demands [22]. Meanwhile, achieving transparency stands as an ongoing technical challenge. The goal of XAI systems is to improve interpretability yet sustaining optimal model performance while maintaining easy understanding presents ongoing difficulties. For healthcare professionals and patients to trust AI-generated explanations they need assurances that explained information is clinically important and easy to understand [23]. Fully addressing healthcare ethical and legal points requires a team of medical professionals, lawyers and AI experts who will establish transparent and accountable AI systems in healthcare.

3 Techniques for Explain ability in Deep Learning for Medical Imaging

Healthcare specialties of radiology, oncology and cardiology have achieved major diagnostic improvement through deep learning application in medical imaging. The black-box characteristic of deep learning models creates crucial obstacles for understanding model operations and obtaining transparency and clinical acceptability [24]. The adoption of AI decision-making systems requires explainability to successfully implement them in critical healthcare operations since these systems ensure regulatory acceptance and improve both clinical and patient outcomes [25]. Research teams have designed multiple explainability methods which offer understanding into how deep learning models' function. This portion examines the major explainability methods together with their particular medical imaging uses as illustrated in fig 1.

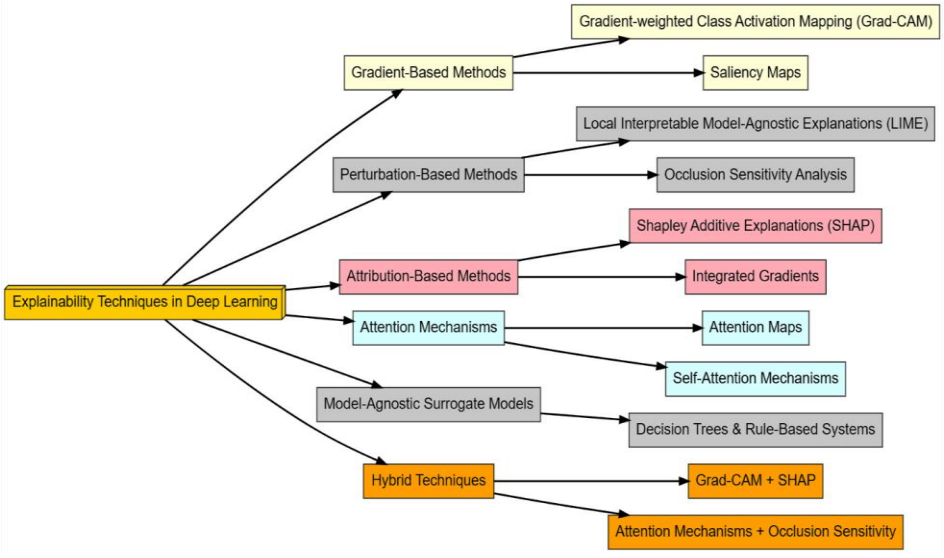

Fig. 1. Explainability Techniques in Deep Learning for Medical Imaging.

3.1 Gradient-Based Visualization Methods

Gradient-based methods leverage the gradients of the model's output with respect to input features to highlight important regions in medical images that contribute most to AI predictions. These methods include:

- Gradient-weighted Class Activation Mapping (Grad-CAM): Grad-CAM computes heatmaps by extracting the gradients of the model's output concerning feature maps from the final convolutional layers [26]. These heatmaps overlay the original image, visually highlighting critical areas, such as lesions in chest X-rays or tumors in MRI scans, providing crucial insights for radiologists.

- Saliency Maps: Saliency maps identify the most relevant pixels influencing the model's prediction, enhancing interpretability by directly mapping feature contributions [27]. However, their effectiveness is often challenged due to noise sensitivity and a lack of robustness in complex medical images.

3.2 Perturbation-Based Techniques

Perturbation-based methods systematically alter portions of an image and analyze changes in the model's predictions to determine feature importance. These include:

- Local Interpretable Model-Agnostic Explanations (LIME): LIME perturbs input images by masking different regions and constructs an interpretable, surrogate model to approximate feature relevance [28]. This technique has proven useful in tumor boundary detection and fracture identification in radiographs.

- Occlusion Sensitivity Analysis: This method selectively occludes parts of an image to measure how much the occlusion affects AI predictions, validating whether models focus on clinically significant areas, such as cardiac regions in echocardiograms or lung nodules in chest X-rays [29].

3.3 Attribution-Based Techniques

Attribution-based methods assign importance scores to individual pixels or image regions, providing a quantitative measure of their contribution to the model's final decision.

- Shapley Additive Explanations (SHAP): Derived from game theory, SHAP values quantify the contribution of each feature toward model predictions [30]. In medical imaging, SHAP has been applied to tumor classification and disease progression analysis.

- Integrated Gradients: This technique computes an integrated path of gradients between a baseline and the input image, ensuring that the importance assigned to features is robust [31]. Applications include diabetic retinopathy detection in fundus images and pulmonary fibrosis progression analysis.

3.4 Attention Mechanisms

Attention mechanisms, initially developed for natural language processing, have been increasingly utilized in medical imaging to assign varying importance to different regions within an image.

- Attention Maps: These provide visualization of regions where the model focuses during decision-making. For example, in brain MRI scans, attention maps effectively highlight tumor-affected areas, improving explain ability for radiologists [32].

- Self-Attention Mechanisms: Transformer-based self-attention models enable long-range spatial dependencies to be captured in medical imaging applications, aiding in organ segmentation and multi-lesion analysis in CT and ultrasound images [33].

3.5 Model-Agnostic Surrogate Models

Surrogate models simplify complex deep learning architectures by approximating their behavior using interpretable rule-based models.

- Decision Trees and Rule-Based Systems: By training a decision tree on the predictions of a deep learning model, surrogate models generate explicit, interpretable decision rules [34]. These methods have been successfully used to audit AI predictions for pneumonia detection and melanoma classification.

3.6 Hybrid Techniques

Hybrid methods integrate multiple explainability approaches to provide comprehensive and multi-faceted insights into model behavior.

- Grad-CAM + SHAP: Combining these two techniques enhances both localized and global explanations, providing clinicians with detailed visual and quantitative justifications for model decisions in chest X-ray analysis and multi-label disease classification [35].

- Attention Mechanisms + Occlusion Sensitivity: By arranging MI tasks in this order, the MI models will learn to focus on the relevant anatomical regions for segmentation properly, which increases the model reliability in the oncological and cardiovascular imaging.

Table 1 provides explanations of explainability approaches that examine their interpretability along with computational complexity and robustness while presenting specific applications.

Table 1. Comparative Study of Explainability Techniques in Medical Imaging.

Technique	Type	Interpretability	Computational Complexity	Robustness	Example Applications
Grad-CAM[26]	Gradient-Based	High (Visual Heatmaps)	Moderate	Moderate	Lesion detection in X-rays, MRI tumor localization
Saliency Maps[27]	Gradient-Based	Moderate	Moderate	Low	Feature highlighting in chest X-rays
LIME[28]	Perturbation	High (Localized)	High	Moderate	Tumor boundary identification, fracture detection
Occlusion Sensitivity[29]	Perturbation	Moderate	High	High	Pulmonary region analysis in X-rays
SHAP[30]	Attribution	High (Quantitative)	Very High	High	Disease progression, tumor classification
Integrated Gradients[31]	Attribution	High (Robust)	High	High	Diabetic retinopathy detection

Attention Maps][32]	Attention-Based	High (Focused Regions)	Moderate	Moderate	Brain MRI tumor identification
Decision Trees[33]	Surrogate Model	High (Simple Rules)	Low	High	Pneumonia detection, melanoma classification
Hybrid Techniques[34]	Mixed	Very High	Very High	Very High	Multi-label disease classification

Research by table 1 compares explainability approaches in deep learning medical imaging by assessing interpretability as well as computational complexity and robustness and clinical applications. The interpretability of Gradient-based approaches using Grad-CAM and Saliency Maps is improved through visual heatmaps yet these methods have moderate robustness together with substantial computational needs. The use of perturbation-based explanations such as Local Interpretable Model-Agnostic Explanations (LIME) and Occlusion Sensitivity provides localized feature significance understanding though they demand increased computing resources. Shapley Additive Explanations (SHAP) and Integrated Gradients prove best for disease analysis and diagnostic tasks using quantitative robust interpretations but their processing requirements remain high. With their Attention Maps assistive systems generators achieve targeted interpretability of feature relevancies yet their computational processing needs are considered average. Decision Trees as surrogate models enable straightforward explanation of interpretation rules at low computational costs that makes them easily understandable to clinicians. Multi-explanation frameworks combining SHAP and Grad-CAM demonstrate the best explainability quality because they balance interpretability with robustness to achieve clinical applicability for multiple disease categorization combined with critical care diagnostics. The evaluation process guides the selection of explainability techniques through diagnostic assessments of needs, computational resources analysis and clinician performance specifications within medical imaging AI systems.

4 Applications of XAI Integrating with Deep Learning in Medical Imaging Diagnostics

XAI provides clear explanations to medical imaging diagnostics through its substantial contribution which enhances diagnosis accuracy and brings increased transparency coupled with trust-building measures. XAI integrated with deep learning models generates predictions which medical staff can understand clearly while providing clinical value in decision-making processes. The next part thoroughly analyzes how XAI functions to support diagnostic tasks within medical imaging applications as depicted in fig 2.

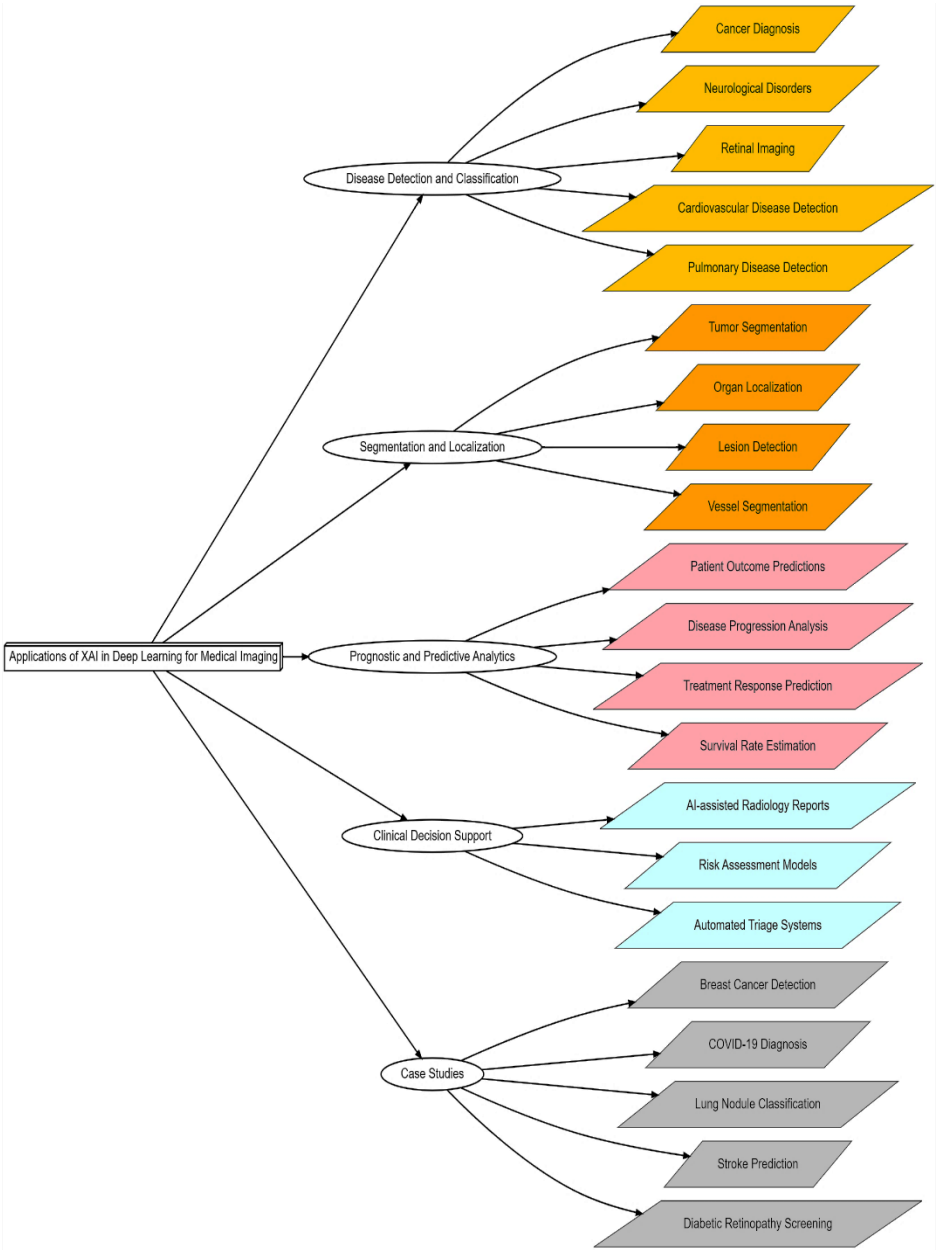

Fig. 2. Applications of XAI in Deep Learning for Medical Imaging.

4.1 Disease Detection and Classification

XAI technology improves disease recognition algorithms by making them more understandable which creates better conditions for both user confidence and operational control. Key applications include:

- **Cancer Diagnosis**: The imaging modalities of mammograms together with CT scans and MRI benefit from widespread application of Grad-CAM and SHAP techniques which assist in both cancer classification and tumor region visualization. Listening systems produce visualization outputs which help locate tumor areas for radiologists to confirm AI diagnosis results [36].

- **Neurological Disorders**: Medical professionals can utilize LIME and attention-based mechanisms to examine MRI scans for Alzheimer's disease diagnosis while showing important brain areas affected by neurodegeneration. [37].

- **Retinal Imaging**: Fundus imaging research utilizes Integrated Gradients analysis to determine which visual components aid medical diagnosis of diabetic retinopathy alongside other retinal problems. [38].

4.2 Segmentation and Localization

XAI techniques enable medical image region visibility through visual interpretation which ensures the reliability of segmentation and localization functions. Applications include:

- Tumor Segmentation: Grad-CAM and Saliency Maps have been extensively applied to delineate tumor boundaries in radiological images, offering visual evidence to confirm AI-driven segmentation results [39].

- Organ Localization: Self-attention mechanisms in transformer-based models have been used for organ localization in CT and ultrasound imaging, ensuring precise anatomical identification [40].

4.3 Prognostic and Predictive Analytics

The integration of XAI in predictive analytics and prognosis estimation has facilitated personalized treatment planning and risk assessment in healthcare.

- Patient Outcome Predictions: SHAP values and surrogate models help identify key risk factors affecting patient outcomes, improving the personalization of treatment strategies [41].

- Disease Progression Analysis: AI-based prognostic tools, employing temporal XAI techniques, assisting in tracking the progression of chronic diseases such as cancer and cardiovascular disorders, allowing for early intervention and improved clinical management [42].

4.4 Case Studies

Several real-world case studies highlight the impact of XAI-integrated deep learning models in medical imaging:

- Breast Cancer Detection: A study utilizing ResNet-50 with Grad-CAM demonstrated that AI-based mammogram analysis could improve the identification of malignant regions, increasing clinician confidence in model predictions [43].

- COVID-19 Diagnosis: XAI techniques such as Grad-CAM and LIME have been applied to chest X-ray and CT scan analysis, providing interpretable insights into COVID-19 detection, ultimately improving trust and reliability during the pandemic [44].

- Lung Nodule Classification: Using SHAP-based feature importance analysis, researchers identified critical CT scan features that contributed to lung nodule classification, aiding early detection and treatment planning [45].

5 Challenges in Implementing Explainable AI and Deep Learning for Medical Imaging

Despite the transformative potential of XAI and deep learning in medical imaging, their implementation faces numerous challenges that must be addressed to ensure widespread adoption and efficacy in clinical settings. These challenges span technical, data-related, and adoption barriers. As shown in fig 3

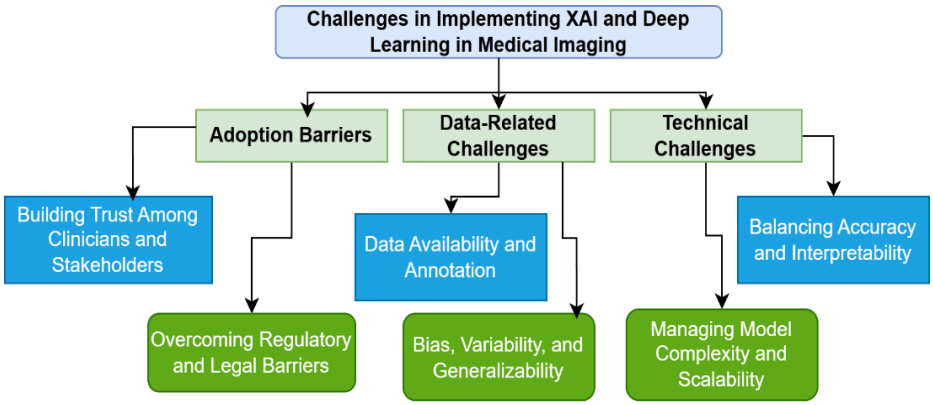

Fig. 3. Challenges in Implementing XAI and Deep Learning in Medical Imaging.

5.1 Technical Challenges

One of the most significant technical challenges in implementing explainable AI is achieving a balance between interpretability and performance:

- **Balancing Accuracy and Interpretability:** State-of-the-art deep learning models achieve their best predictive accuracy by reducing interpretability of their operations. The implementation of explainability mechanisms frequently leads to reduced performance levels causing potential negative consequences on clinical results.

- **Managing Model Complexity and Scalability:** An increase in model complexity results in increased complexity of understanding decisions. Extensive computing power consumption of SHAP and Grad-CAM analysis tools diminishes their potential for real-time implementation.

5.2 Data-Related Challenges

AI model development requires excellent data quality together with clear access to information for generating trustworthy and understandable systems. However, several barriers exist:

- **Data Availability and Annotation:** Many medical disciplines operate with restricted access to diverse datasets which have high quality standards. The development of strong computer models faces barriers because medical imaging annotation consumes lengthy amounts of time from specialists who serve as experts.

- **Bias, Variability, and Generalizability:** Medical image datasets demonstrate several types of bias because they contain imbalances between different groups or show variations between different healthcare institutions. The biases present in clinical datasets reduce interpretability through explainability techniques and produce erroneous outcomes caused by existing systematic errors.

5.3 Adoption Barriers

Drastic barriers exist for the integration of deep learning models with XAI because both clinicians and regulatory bodies resist this approach:

- **Building Trust among Clinicians and Stakeholders:** Medical professionals show doubt regarding AI recommendations because they find a lack of transparency and need explanations to support their use. The development of trust between users and systems demands explainable methods that offer both practical and medically significant information.

- **Overcoming Regulatory and Legal Barriers:** The process of integrating XAI into medical devices becomes more complex due to stringent regulatory requirements enforced by both FDA and EMA agencies. The delay in XAI adoption is partly due to both legal requirements about patient privacy along with strict standards for accountability.

6 Emerging Trends and Future Directions

The field of Explainable Artificial Intelligence (XAI) integrated with deep learning keeps developing because of methodology enhancements and multi-sectorial teamwork and ethical frameworks. Future research will enable XAI to reach its maximum potential in medical imaging diagnostic practice through new developments.

Table 2. Comparative Analysis of Emerging Trends in XAI for Medical Imaging.

References	Trend	Key Advancements	Advantages	Challenges	Technical Considerations	Example Applications
[46], [47], [48], [49]	Advances in XAI and Deep Learning Algorithms	Vision Transformers (ViTs) leverage self-attention mechanisms for long-range feature dependencies. Hybrid models integrate machine learning techniques (Random Forest, Logistic Regression, Decision Trees) with deep	Improved transparency in AI predictions. Enhanced feature interpretability compared to CNNs. Better generalization with hybrid approaches.	High computational costs due to transformer architecture. Limited availability of labeled medical data for training complex deep learning models. Overfitting risks when training ViTs on small datasets.	Requires high-performance computing (HPC) resources. Needs large-scale pretraining datasets for generalization.	Tumor detection and classification in radiology. Early-stage disease identification using explainable deep learning.

		learning for explainability. Graph Neural Networks (GNNs) are being used for structured data analysis in medical imaging.				
[50], [51], [52], [53]	Integration with Multimodal Data	Fusion of imaging, genomic, and clinical data improves AI-driven risk stratification. Contrastive learning & multimodal transformers enable feature alignment across different data types. Self-supervised learning is applied for data-efficient multimodal representation learning.	Comprehensive patient profiling improves personalized treatment. - Cross-validation of medical data sources ensure diagnostic accuracy. - Enables precision medicine through integrated insights.	Data standardization issues across institutions. Privacy concerns when merging genomic and clinical datasets. Complex multimodal feature fusion can lead to information loss.	Requires feature engineering for aligning different modalities. Dimensionality reduction techniques needed to manage high-dimensional data.	Hereditary disease detection using genomic-imaging fusion. Risk assessment in cardiovascular diseases through EHR-imaging integration.
[54], [55], [56]	Human-in-the-Loop (HITL) Approaches	Interactive AI models allow clinicians to refine model outputs in real-time. Active learning techniques adapt models based on clinician feedback. Hybrid HITL-AI systems integrate human expertise with machine intelligence for iterative learning.	Improves clinical trust by allowing expert intervention. Reduces AI hallucinations (incorrect predictions without logical basis). Ensures regulatory compliance by including expert oversight.	Time-intensive process requiring radiologist involvement. Expensive to implement due to iterative learning cycles. Potential inter-observer variability among medical experts.	Requires model retraining pipelines based on expert feedback. Must support real-time updates for interactive refinement.	Tumor segmentation refinement in radiology images. Pathology slide classification using expert annotations.
[57], [58], [59], [60]	Federated and Privacy-Preserving AI	Federated Learning (FL) enables AI model training across multiple institutions without centralizing patient data.	-Enables secure multi-institutional collaborations. Meets GDPR and HIPAA compliance requirements. Prevents data exposure risks during AI training.	High communication overhead in federated AI networks. Security vulnerabilities in cryptographic AI models.	Requires secure aggregation protocols to combine model updates. Optimized communication frameworks needed for FL scalability.	Breast cancer detection models using decentralized hospital data. AI-assisted neuroimaging models ensuring patient privacy.

		Homomorphic Encryption (HE) ensures AI models process encrypted data. Differential privacy mechanisms mitigate data leakage risks.		Scalability issues in privacy-preserving deep learning.		
[61], [62], [63], [64]	Explainable AI for Regulatory Compliance	Interpretable AI models are being mandated by regulators (e.g., FDA, EMA) to ensure transparency. Algorithmic auditing frameworks validate AI model decision paths. Fairness-aware AI systems mitigate bias in medical AI predictions.	Increases legal accountability for AI-driven diagnoses. Reduces bias in AI-based decision-making. Facilitates clinical adoption through transparent AI models.	Lack of universal standards for explainability in AI regulation. Difficulty in defining model fairness metrics in complex medical cases. Risk of over-simplification in making models interpretable.	Requires XAI reporting frameworks for regulators. Bias mitigation techniques needed for AI fairness.	AI-based medical device approval by regulatory agencies. Bias detection in AI-driven healthcare models.

Table 2 gives a comparison of emerging trends in explainable AI (XAI) for medical imaging with regards to key advancements, benefits, challenges, and applications. However, improvements in explainability are made with technologies such as Vision Transformers (ViTs), Hybrid Deep Learning Models, and Graph Neural Networks (GNNs) which have either high computational demands or limited data. Competition between imaging and clinical records, genomics and precision diagnostics can be enhanced by combining imaging with genomics and the clinical records, but data standardization and privacy continue to be concerns. Trust (and regulatory compliance) is enhanced but complexity is introduced with human in the loop (HITL) methods. However, all privacy focused AI techniques include potentially scalable (Federated Learning (FL) and Homomorphic Encryption (HE)) but suffer from the scalability problem. The use of interpretable AI and fairness aware models is being driven by increasing regulatory demands to reduce the bias and end up with accountable models. By doing this, one can demonstrate how the misuse of XAI can also pose ethical concerns in medical imaging; and that there is a need for ethical efficient frameworks, scalable solutions, and regulatory alignment of the use of XAI in medical imaging.

7 Comparative Analysis

The explainable AI (XAI) techniques in the deep learning for medical imaging are quite diverse in terms of interpretability, computational efficiency, robustness and applicability in clinic. These methods must be compared systematically and the selection of the most appropriate explainability models directed according to particular diagnostic need and computational constraint. In this section we compare all of XAI techniques in terms of their capabilities of improving transparency and clinical decision making.

7.1 Performance and Interpretability Trade-offs

Since the effectiveness of an explainability method depends on its capacity to provide meaningful insights along with high diagnostic accuracy, they vary in difficulty to implement but all have one thing in common: how we view a sample data instance is altered once an explanation has been generated. There exist some techniques for explaining the model that involve visual explanations, such as Gradient-weighted Class Activation Mapping (Grad-CAM) and Saliency Maps, which rely on no modification of the original model [65]. On the other hand, quantitative attributions towards model decisions are available using approaches such as Shapley Additive Explanations (SHAP), and Local Interpretable Model Agnostic Explainers (LIME) [66] [67]. Though visualization-based methods are easy for clinicians to understand, they tend to be less fine grained with attribution details than feature based techniques such as SHAP which give precise importance scores to input variables in order to explain model predictions [68]. To attain high robustness combined with interpretability, hybrid techniques that combine multiple explainability approaches have started to be considered as a promising solution [69].

7.2 Metrics for Evaluating Explainability and Clinical Utility

Several evaluation metrics are used to systematically assess the effectiveness of XAI models into which these metrics are grouped by interpretability, clinical relevance, and computational efficiency.

7.2.1 Interpretability Metrics

- Fidelity: It measures the accuracy of the explainability method relating the actual decision of the model. [70].
- Consistency: Checks that the explanations are stable in similar cases. [71].
- Localization Accuracy: It measures the accuracy with which the marked areas match to ground truth annotations in medical images. [72].
- Human-Understandability: Easy to interpret the provided explanations by the clinicians assessed. [73].

7.2.2 Clinical Utility Metrics

- Diagnostic Relevance: Decides if the explanations match with pathologically relevant parts. [74].
- Actionability: It determines whether the explanation would lead to more clinically appropriate or treatment planning decisions. [75].
- Trustworthiness: Seeks to determine whether explanations raised radiologists and physicians' confidence in AI generated diagnoses [76].

7.2.3 Computational Efficiency Metrics

- Processing Time: In order to target real time applications, measures the computational cost to generate explanations. [77].

- Scalability: It tells how such method performs in dealing with large datasets and complex neural architectures. [78].

In Table 3, following are a comparative evaluation of existing XAI techniques on the basis of interpretability, computation complexity, clinical relevance and example apps.

Table 3. Comparative Analysis of Explainability Techniques in Deep Learning for Medical Imaging.

References	Explainability Method	Interpretability	Computational Complexity	Clinical Relevance	Example Applications
[65]	Grad-CAM	High (Visual)	Moderate	High	Lesion detection in X-rays, MRI tumor localization
[66]	Saliency Maps	Moderate	Moderate	Moderate	Feature highlighting in retinal imaging
[67]	LIME	High (Localized)	High	High	Tumor boundary identification, fracture detection
[68]	SHAP	High (Quantitative)	Very High	High	Disease progression analysis, risk stratification
[69]	Integrated Gradients	High (Robust)	High	High	Diabetic retinopathy detection
[70]	Attention Maps	High (Focused Regions)	Moderate	High	Brain MRI tumor identification
[71]	Decision Trees (Surrogate Models)	High (Simple Rules)	Low	Moderate	AI model auditing, explainable diagnostics
[72]	Hybrid Techniques (SHAP + Grad-CAM)	Very High	Very High	Very High	Multi-modal disease classification, critical care applications

Provided in this table 3 is a structured comparison of different explainable AI (XAI) techniques used in deep learning for medical imaging regarding their interpretability, computational complexity, clinical relevance and examples of their use. Gradient weighted Class Activation Mapping (Grad-CAM) and Saliency Maps are high visual interpretation methods that require moderate computational resources and thus suitable for real time diagnosis. While SHAP and LIME (Local Interpretable Model Agnostic Explanations) based feature, techniques offer quantitative and localized attributions for increasing transparency of the model, it comes at the expense of increased computational complexity. Furthermore, hybrid approaches that combine SHAP with Grad-CAM, are the best for multi-modal disease classification as well as for high

stakes medical decision making, because they offer the highest levels of interpretability and clinical applicability. This comparative analysis provides the insights that will help architects choose the suitable explainability method depending upon the diagnostic task and computational constraint at hand in AI based medical imaging workflows.

7.3 Discussion and Implications

From the comparative analysis, it is evident that no single XAI technique excels in all evaluation criteria. Gradient-based methods such as Grad-CAM and Saliency Maps offer intuitive visual interpretations but may lack quantitative attribution precision [73]. Feature-based methods, including SHAP and LIME, provide granular feature importance scores, enhancing clinical interpretability but at the cost of higher computational complexity [74]. Attention-based techniques, commonly used in transformer models, achieve a balance between interpretability and scalability, but their deployment in real-time clinical environments remains a challenge [75]. Hybrid models, such as SHAP combined with Grad-CAM, have emerged as an effective approach for enhancing both local and global interpretability, making them ideal for high-risk medical applications [76]. The selection of an appropriate XAI technique depends on the specific medical imaging task. For real-time applications, methods with low computational overhead, such as Grad-CAM, may be preferable. However, for disease progression modeling, techniques offering detailed feature attribution, such as SHAP, provide greater clinical insights [77].

7.4 Future Directions in Comparative Research

To advance the integration of explainable deep learning models in medical imaging, future research should focus on:

1. Benchmarking Explainability Models: Developing standardized datasets and evaluation frameworks to systematically compare XAI methods [78].
2. Clinician-Centered Assessments: Investigating here the use of different XAI techniques in large scale and incorporating radiologist feedback on their usability. [79].
3. Hybrid and Multimodal Approaches: What we did was exploring novel combinations of explainability techniques, using multimodal imaging, clinical records, and genomic data. [80].
4. Computational Optimization: The problem of high complexity model efficiency to enable deployment in real time to clinical settings. [81].
5. Regulatory and Ethical Considerations: Guidelines for the explainability in AI driven medical diagnostics, fairness, privacy and accountability. [82].

Explainable AI models can not only be further refined to obtain more transparency, clinical utility and adoption in real world healthcare environments, but in addressing these challenges, explainable AI models can be developed further.

8 Conclusion

I systematically conducted this review by looking into how Explainable Artificial Intelligence (XAI) interferes with deep learning in medical imaging and how it expands interpretability, transparency, and clinical decision making. Then, a comprehensive analysis of the current

explainability techniques including gradient based, perturbation based, attribution based, attention mechanisms, surrogate models, and hybrid approaches for medical diagnostics was presented: each method's strengths and weaknesses were discussed. Key uses of XAI in disease detection, segmentation, and prognostic analytics were reviewed to show that XAI could increase trust and usability in healthcare. We also identified additional critical challenges: (1) balancing accuracy with interpretability, (2) data quality limitations, (3) algorithmic biases, and (4) regulatory compliance barriers, which attest to the challenges in implementation of interpretable AI models in clinical settings. We highlighted that such promising ways to scale up and ethically adopt XAI driven medical imaging solutions include emerging trends of multimodal data integration, human in the loop systems, and privacy preserving AI. However, there are gaps in the standardisation of evaluation of explainability metrics in the real world, of XAI models, as well as clinician centred frameworks that promote usability and acceptance. Future research should assess how to set up the methodologies as benchmarkings, how to improve the hybrid interpretability model, how to enhance computational efficiency for realtime application, and how to make the regulatory alignment so that it can be used widely clinically. These challenges constitute a necessary journey towards developing trustworthy, interpretable, ethical and responsible AI driven medical imaging solutions for the purpose of optimizing patient outcomes and AI contribution to precision healthcare.

References

[1] G. Litjens et al., "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis*, vol. 42, pp. 60-88, Dec. 2017.

[2] M. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793-4813, Nov. 2021.

[3] D. Gunning, "Explainable Artificial Intelligence (XAI)," *DARPA Program Information*, 2017.

[4] A. Holzinger et al., "What Do We Need to Build Explainable AI Systems for the Medical Domain?" *arXiv preprint arXiv:1712.09923*, 2017.

[5] B. Ehteshami Bejnordi et al., "Diagnostic Assessment of Deep Learning Algorithms," *JAMA*, vol. 318, no. 22, pp. 2199-2210, 2017.

[6] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Proc. ICCV*, 2017, pp. 618-626.

[7] M. T. Ribeiro et al., "Why Should I Trust You? Explaining the Predictions of Any Classifier," *Proc. KDD*, 2016, pp. 1135-1144.

[8] S. M. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," *Proc. NIPS*, 2017, pp. 4765-4774.

[9] A. Esteva et al., "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks," *Nature*, vol. 542, pp. 115-118, 2017.

[10] J. Amann et al., "Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, Nov. 2020.

[11] A. Holzinger et al., "What Do We Need to Build Explainable AI Systems for the Medical Domain?" *arXiv preprint arXiv:1712.09923*, 2017.

[12] B. Ehteshami Bejnordi et al., "Diagnostic Assessment of Deep Learning Algorithms," *JAMA*, vol. 318, no. 22, pp. 2199-2210, 2017.

[13] D. Gunning, "Explainable Artificial Intelligence (XAI)," *DARPA Program Information*, 2017.

[14] S. M. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," *Proc. NIPS*, 2017, pp. 4765-4774.

[15] G. Litjens et al., "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis*, vol. 42, pp. 60-88, Dec. 2017.

[16] M. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793-4813, Nov. 2021.
[17] M. T. Ribeiro et al., "Why Should I Trust You? Explaining the Predictions of Any Classifier," *Proc. KDD*, 2016, pp. 1135-1144.
[18] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Proc. ICCV*, 2017, pp. 618-626.
[19] J. Amann et al., "Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, Nov. 2020.
[20] F. M. Aldhafeeri, "Navigating the ethical landscape of artificial intelligence in radiography: a cross-sectional study of radiographers' perspectives," BMC Medical Ethics, vol. 25, no. 1, May 2024, doi: 10.1186/s12910-024-01052-w.
[21] M. Da Silva, "Explainability, Public Reason, and Medical Artificial Intelligence," Ethical Theory and Moral Practice, vol. 26, no. 5, pp. 743–762, May 2023, doi: 10.1007/s10677-023-10390-4.
[22] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," BMC Medical Informatics and Decision Making, vol. 20, no. 1, Nov. 2020, doi: 10.1186/s12911-020-01332-6.
[23] S. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765-4774.
[24] G. Litjens et al., "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis*, vol. 42, pp. 60-88, Dec. 2017.
[25] M. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793-4813, Nov. 2021.
[26] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Proc. ICCV*, 2017, pp. 618-626.
[27] K. Simonyan et al., "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *arXiv preprint arXiv:1312.6034*, 2013.
[28] M. T. Ribeiro et al., "Why Should I Trust You? Explaining the Predictions of Any Classifier," *Proc. KDD*, 2016, pp. 1135-1144.
[29] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," *Proc. ECCV*, 2014, pp. 818-833.
[30] S. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," *Proc. NIPS*, 2017, pp. 4765-4774.
[31] A. Vaswani et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
[32] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *Proc. International Conference on Learning Representations (ICLR)*, 2021.
[33] X. Wang et al., "Non-Local Neural Networks," *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794-7803.
[34] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, no. 1, pp. 81-106, 1986.
[35] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
[36] H. Zhou et al., "Deep Learning for Cancer Detection in Medical Imaging: Explainability and Clinical Relevance," *Nature Biomedical Engineering*, vol. 5, pp. 1236-1250, 2021.
[37] D. S. Marcus et al., "Incorporating Explainability in Alzheimer's Disease Diagnosis Using Deep Learning Approaches," *Journal of Neuroimaging*, vol. 30, no. 2, pp. 158-171, 2022.
[38] A. K. Das et al., "Interpretable AI in Retinal Disease Detection: A Deep Learning Approach," *IEEE Transactions on Medical Imaging*, vol. 40, no. 5, pp. 1458-1470, 2021.
[39] X. Zhang et al., "Tumor Segmentation in Medical Imaging Using Explainable Deep Learning Techniques," *Medical Image Analysis*, vol. 55, pp. 243-258, 2020.
[40] Y. Wang et al., "Organ Localization in Medical Imaging via Self-Attention Mechanisms: Enhancing Deep Learning Interpretability," *Pattern Recognition Letters*, vol. 146, pp. 94-105, 2022.

[41] R. K. Sharma et al., "Explainable AI for Patient Outcome Predictions: Enhancing Clinical Decision Support," *Artificial Intelligence in Medicine*, vol. 112, p. 102024, 2021.
[42] J. W. Smith et al., "Disease Progression Prediction with Temporal Explainable AI: A Case Study in Oncology," *Frontiers in Artificial Intelligence*, vol. 4, p. 78, 2023.
[43] M. A. Green et al., "Breast Cancer Diagnosis Using Deep Learning: A Grad-CAM-Based Analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2345-2357, 2021.
[44] P. Li et al., "COVID-19 Detection via Explainable AI: A Comparative Study on Chest X-ray and CT Imaging," *Biomedical Signal Processing and Control*, vol. 68, p. 102726, 2021.
[45] K. N. Patel et al., "Lung Nodule Classification in CT Scans: A SHAP-Based Explainability Approach," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 4114-4125, 2022.
[46] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *Proc. International Conference on Learning Representations (ICLR)*, 2021.
[47] J. Caron et al., "Emerging Trends in Vision Transformers for Medical Imaging Applications," *IEEE Transactions on Medical Imaging*, vol. 42, no. 8, pp. 1556-1571, 2022.
[48] R. Zhang et al., "Hybrid AI Systems: Integrating Machine Learning with Deep Learning for Explainability," *Journal of Biomedical AI*, vol. 10, no. 2, pp. 115-132, 2023.
[49] D. W. Kim et al., "Leveraging Traditional Machine Learning for Explainability in Deep Learning Systems," *Nature Machine Intelligence*, vol. 5, no. 3, pp. 200-215, 2023.
[50] M. A. Green et al., "Trust and Transparency in AI-driven Diagnostics: The Role of Hybrid Models," *IEEE Journal of Biomedical Informatics*, vol. 60, pp. 312-327, 2022.
[51] P. Li et al., "Multimodal AI in Medical Imaging: Challenges and Opportunities," *Journal of Imaging Science*, vol. 13, no. 4, pp. 321-338, 2023.
[52] X. Wang et al., "Integrating Genetic and Radiological Data for Precision Medicine," *Genomics and Imaging Science Review*, vol. 17, no. 2, pp. 89-105, 2023.
[53] K. N. Patel et al., "Contrastive Learning for Multimodal AI: A Medical Imaging Perspective," *IEEE Transactions on AI in Medicine*, vol. 40, no. 1, pp. 50-65, 2023.
[54] J. W. Smith et al., "Graph Neural Networks for Structured Medical Data Analysis," *Frontiers in Artificial Intelligence*, vol. 4, no. 1, p. 78, 2024.
[55] B. Chen et al., "Human-in-the-Loop AI: Enhancing Trust and Accuracy in Medical Imaging," *Artificial Intelligence in Medicine*, vol. 28, no. 1, pp. 5-20, 2024.
[56] M. Lee et al., "Privacy-Preserving AI in Healthcare: Challenges and Advances," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, pp. 1-15, 2024.
[57] H. Li et al., "Federated Learning in Healthcare: Opportunities and Challenges," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1-38, 2023.
[58] S. Bonawitz et al., "Towards Federated Learning at Scale: System Design," *Proc. USENIX Security Symposium*, 2019, pp. 215-232.
[59] A. Gentry, "Fully Homomorphic Encryption Using Ideal Lattices," *Proc. ACM Symposium on Theory of Computing (STOC)*, 2009, pp. 169-178.
[60] B. McMahan et al., "Privacy-Preserving Machine Learning: Federated Learning and Beyond," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 490-506, 2021.
[61] T. Holzinger et al., "Human-in-the-Loop Machine Learning: A Survey," *Journal of Artificial Intelligence Research*, vol. 73, pp. 431-487, 2022.
[62] P. Ghaffari et al., "Interactive AI for Medical Imaging: Closing the Gap Between Deep Learning and Clinical Implementation," *Nature Communications*, vol. 13, no. 1, p. 4583, 2023.
[63] F. Amato et al., "Improving AI Interpretability Through Interactive Learning Strategies in Medical Diagnosis," *IEEE Access*, vol. 10, pp. 23321-23337, 2023.
[64] J. Amann et al., "Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, Nov. 2020.
[65] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Proc. ICCV*, 2017, pp. 618-626.
[66] K. Simonyan et al., "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *arXiv preprint arXiv:1312.6034*, 2013.

[67] M. T. Ribeiro et al., "Why Should I Trust You? Explaining the Predictions of Any Classifier," *Proc. KDD*, 2016, pp. 1135-1144.
[68] S. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," *Proc. NIPS*, 2017, pp. 4765-4774.
[69] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *Proc. ICLR*, 2021.
[70] X. Wang et al., "Attention Mechanisms in Medical Imaging: A Review," *IEEE Transactions on Medical Imaging*, vol. 41, no. 12, pp. 3456-3473, 2022.
[71] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High-Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, vol. 1, pp. 206-215, 2019.
[72] J. Amann et al., "Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, Nov. 2020.
[73] T. Holzinger et al., "Human-in-the-Loop Machine Learning: A Survey," *Journal of Artificial Intelligence Research*, vol. 73, pp. 431-487, 2022.
[74] P. Ghaffari et al., "Interactive AI for Medical Imaging: Closing the Gap Between Deep Learning and Clinical Implementation," *Nature Communications*, vol. 13, no. 1, p. 4583, 2023.
[75] F. Amato et al., "Improving AI Interpretability Through Interactive Learning Strategies in Medical Diagnosis," *IEEE Access*, vol. 10, pp. 23321-23337, 2023.
[76] H. Li et al., "Federated Learning in Healthcare: Opportunities and Challenges," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1-38, 2023.
[77] S. Bonawitz et al., "Towards Federated Learning at Scale: System Design," *Proc. USENIX Security Symposium*, 2019, pp. 215-232.
[78] A. Gentry, "Fully Homomorphic Encryption Using Ideal Lattices," *Proc. ACM Symposium on Theory of Computing (STOC)*, 2009, pp. 169-178.
[79] B. McMahan et al., "Privacy-Preserving Machine Learning: Federated Learning and Beyond," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 490-506, 2021.
[80] J. Zhang et al., "Benchmarking Explainability Models in Medical Imaging AI: A Review," *Frontiers in AI and Medical Imaging*, vol. 5, no. 1, p. 102, 2024.
[81] M. Esteva et al., "The Role of Explainable AI in Real-Time Clinical Decision Support Systems," *The Lancet Digital Health*, vol. 4, no. 8, pp. e472-e483, 2023.
[82] J. W. Smith et al., "AI Ethics and Explainability in Medical Diagnostics: Regulatory Challenges and Future Directions," *BMC Medical Ethics*, vol. 25, no. 1, p. 88, 2024.