



(REVIEW ARTICLE)



TRUSTWORTHY ARTIFICIAL INTELLIGENCE FOR MEDICAL IMAGE ANALYSIS: A COMPREHENSIVE REVIEW OF ROBUST DEEP LEARNING TECHNIQUES FOR CLINICAL DECISION SUPPORT

Rebecca Yaa Bonah ^{1,*} and Ebenezer Tetteh ²

¹ University of South Dakota, Vermillion, USA

² Department of Statistics and Actuarial Science, University of Ghana

* Corresponding Author

Magna Scientia Advanced Research and Review, 2026, 18(01), 182–196

Publication history: Received on 21 August 2026; revised on 29 September 2026; accepted on 01 October 2026

Article DOI: <https://doi.org/10.30574/msarr.2026.18.1.0196>

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

Medical image analysis has been revolutionized by the advent of artificial intelligence, which has made automated disease detection, segmentation and diagnosis possible across a range of imaging modalities, thus supporting clinical decisions. However, the increasing complexity of deep learning models has raised concerns regarding their reliability, interpretability, fairness, privacy, and regulatory compliance, highlighting the need for trustworthy AI in healthcare. The review focuses on recent breakthroughs in deep learning models for medical image analysis, such as Convolutional Neural Networks (CNNs), segmentation networks based on U-net, Vision Transformers, and foundation models, and explores their use in clinical decision support. It also examines the fundamental principles of trustworthy AI, including explainability, robustness, fairness, privacy preservation, reliability, transparency and regulatory governance. The review summarizes the current state of the art for the strengths and limitations of current AI models and highlights some of the challenges that hamper the use of these models in the clinical workflow, such as limited data availability, domain shifts, algorithmic bias, computational complexity, and integration into the clinical workflow. New concepts like multimodal foundation models, federated learning and privacy-preserving AI are also explored as potential advancements in the safe, scalable and clinically relevant use of AI in medical imaging. The review highlights the significant progress medical image analysis has made using deep learning techniques but emphasizes that the effective translation of these technologies into clinical practice requires both technical advancement and adherence to trustworthy AI principles and effective governance models. To ensure these AI systems are effective for patient care and to aid in evidence-based medical decision making, development needs continued cooperation between researchers, clinicians, healthcare institutions and regulatory bodies, with the aim of creating reliable, transparent and clinically acceptable AI systems.

Keywords: Trustworthy Artificial Intelligence, Medical Image Analysis, Deep Learning, Clinical Decision Support, Foundation Models

1 INTRODUCTION

Medical image analysis is one of the most successful applications of AI in modern healthcare, transforming the way medical images are analyzed and interpreted. In modern healthcare, medical image analysis is a successful application

of AI that is reshaping the way medical images are analyzed and interpreted. With the explosion of digital imaging technologies, there have been vast amounts of medical data generated from magnetic resonance imaging (MRI), computed tomography (CT), X-ray, ultrasound, histopathology, retinal imaging, and endoscopy. With growing imaging workloads, complexity of disease and inter-observer variation, it is still a challenging task for clinicians to interpret these images correctly and efficiently. In recent years, the process has significantly been enhanced by deep learning, which learns hierarchical features of images directly from the raw data, resulting in highly accurate image classification, lesion detection, segmentation, disease diagnosis and prognosis prediction in a wide variety of clinical applications (Rajpurkar et al., 2022; Litjens et al., 2017).

Advances in neural network architecture have been the key to the success of deep learning in medical imaging. The introduction of Convolutional Neural Networks (CNNs) marked the beginning of automated medical image analysis that was able to replace handcrafted feature extraction, and more recent advances such as Vision Transformers (ViTs), transformer-based segmentation networks, self-supervised learning and multimodal foundation models have further enhanced the representation of features, the capacity for context and effective model generalization (Litjens et al., 2017; Dosovitskiy et al., 2020; Hatamizadeh et al., 2022; Chen et al., 2026). These models have shown excellent results across various domains like radiology, dermatology, pathology, ophthalmology, and oncology, aiding physicians in making more accurate diagnoses, minimizing interpretation times, and making informed decisions (Hartsock et al., 2026; Esteva et al., 2017; Topol, 2019).

These successes fall short of requirements for clinical use, however, because predictive accuracy is not enough. Most deep learning models operate as ‘black box’ systems, meaning that their decision-making process is hard to interpret for clinicians and thus less people are confident in the AI-assisted diagnosis process. Furthermore, the predictive performance of the model may degrade when deploying data from other hospitals, scanners, or patient populations due to adversarial attacks, imaging artifacts and distribution shifts. Further issues touch upon the demographic bias, lack of transparency in the models, prediction uncertainty, privacy concerns of sensitive patient information, and the lack of standardized governance mechanism for safe deployment (Amann et al., 2020; Holzinger et al., 2019; Apostolidis & Papakostas, 2021; Zhou et al., 2022; Ueda et al., 2024; Zhu et al., 2024).

Reliability, explainability, robustness, fairness, transparency, privacy preservation, accountability, regulatory compliance, and predictive performance are all important aspects of trustworthy AI and have become a focus of research, in contrast to creating AI that performs accurately. International organizations and regulatory agencies, including the World Health Organization (WHO), the United States Food and Drug Administration (FDA), the National Institute of Standards and Technology (NIST), and the European Union, have introduced governance frameworks that emphasize responsible AI development, continuous monitoring, human oversight, and risk management throughout the AI lifecycle (WHO, 2024; Joshi et al., 2024; NIST, 2023; Act, 2024). Meanwhile, new technologies like foundation models, vision-language models, segment anything (SAM), biomedical (BiMed) CLIP, and LLaVA-Med are adding new challenges in terms of explainability, computational complexity, clinical validation, and governance, while simultaneously expanding the capabilities of medical image analysis (Chen et al., 2026; Kirillov et al., 2023; Ma et al., 2024; Zhang et al., 2023; Li et al., 2023).

While there have been many studies that have looked at deep learning systems, explainable AI, foundation models, or AI governance individually, few studies have taken a broad view of the impact these advances can have on trustworthy medical image analysis and clinical decision support. This review fills that void by summarizing the latest evidence on robust deep learning methods and assessing them using the lens of trustworthy AI. It discusses current deep learning architectures and the use of explainability, robustness, fairness, privacy, transparency, and regulatory frameworks in medical imaging, summarizes the pros and cons of these, and investigates future trends and research directions for the development of reliable AI-assisted medical imaging systems.

2 ARTIFICIAL INTELLIGENCE IN MEDICAL IMAGE ANALYSIS

Artificial Intelligence has revolutionized the field of medical image analysis, making the task of analyzing complex medical images automated, faster, more consistent and accurate. Historically, image interpretation was a completely manual process and depended solely on the knowledge and experience of radiologists, pathologists and other specialists, which led to inter-observer variability, fatigue and reduced time to specialize. These challenges are exacerbated by the acceleration of digital imaging technologies and the generation of even more imaging data that must be processed quickly and accurately. These challenges are overcome using AI for image classification, lesion detection, segmentation, diagnosis, and prognosis, which enhances the efficiency of diagnosis and aids clinical decisions based on evidence (Rajpurkar et al., 2022; Hartsock et al., 2026).

As AI has developed, medical imaging has shifted from traditional methods that primarily relied on handcrafted features to deep learning models that are able to automatically extract complex representations from medical imaging data. This

progress has greatly enhanced the diagnostic performance of many clinical specialties and imaging modalities and presented new opportunities and challenges regarding model reliability, interpretability, robustness, fairness and clinical deployment. Therefore, it is critical to understand the characteristics of medical imaging modalities, the evolution of AI techniques, and the use of AI in clinical decision support systems. These will be explored more in-depth in the subsections below.

2.1 Medical Imaging Modalities

Medical imaging is the cornerstone of the modern clinical diagnosis, enabling the non-invasive imaging of anatomical structures, physiological processes and pathological abnormalities. The variety of imaging modalities allows physicians to diagnose and track a wide variety of diseases with greater accuracy, and the resulting large amount of complex data opens the door to using AI to help analyze it. Different imaging modalities have unique physical properties, image quality, spatial resolution and diagnostic uses, which means that AI models must be modality-specific to address the unique challenges each imaging modality presents. Although numerous medical imaging modalities are used in clinical practice, this review focuses on the principal modalities that have driven recent advances in AI-based medical image analysis, namely Magnetic Resonance Imaging (MRI), Computed Tomography (CT), X-ray imaging, ultrasound, digital pathology, ophthalmic imaging, and endoscopic imaging. These modalities were selected because they represent the most widely studied and clinically relevant imaging domains in the literature reviewed and illustrate the breadth of AI applications across diverse diagnostic settings. Each modality possesses distinct physical characteristics, image quality, spatial resolution, and diagnostic requirements, necessitating AI models that can address modality-specific challenges. Therefore, it is important to consider the nature of the imaging data that the deep learning algorithm needs to process when selecting and evaluating the algorithm (Litjens et al., 2017; Hartsock et al., 2026).

MRI is a technique which is commonly used for imaging soft tissues due to its superior contrast resolution and the fact that it does not employ any ionizing radiation. MRI is a significant tool used in diagnosis of neurological disorders, musculoskeletal injuries, cardiovascular diseases and various cancers. The MRI datasets are, however, affected by variability such as scanner, acquisition protocol, imaging sequence and patient motion, and automated image interpretation is especially difficult. AI has shown great promise in supporting MRI analysis, including the detection of tumors, segmentation of tissues, classification of diseases, reconstruction of images, and quantitative measurements, all of which help to enhance the efficiency and precision of MRI diagnosis (Litjens et al., 2017; Hartsock et al., 2026).

Computed Tomography (CT) is a technique that generates detailed images of cross-sections with the help of X-rays and is essential in emergency medicine, oncology, trauma diagnostics and cardiovascular imaging. CT imaging provides very good visualization of internal organs, bones and vascular structures, and can quickly detect abnormalities like hemorrhage, fractures, pulmonary diseases and malignant lesions. Some examples of automated CT image processing tasks using AI are organ segmentation, lesion detection, disease classification, image reconstruction and radiation dose optimization. However, the fact that the scanners and reconstruction algorithms are different and the acquisition settings vary can lead to distribution shifts that impact the robustness and generalizability of the AI models across healthcare institutions (Hartsock et al., 2026; Su et al., 2024).

X-ray imaging is the most used form of diagnostic imaging since it is fast, readily available and relatively cheap. Chest radiography is especially important in the diagnosis of respiratory diseases, such as pneumonia, tuberculosis, lung cancer, fractures, and cardiovascular diseases. The performance of deep learning algorithms in automated chest X-ray interpretation is outstanding, with the ability to recognize subtle imaging patterns that might be challenging to identify in a consistent manner during routine clinical work. While these developments are promising, persistent issues like image overlap, low contrast resolution, and demographic bias still pose challenges that necessitate the development of strong and equitable AI models that can deliver consistent performance across a wide range of patient populations (Hartsock et al., 2026; Seyyed-Kalantari et al., 2021).

Ultrasound imaging gives the physician a real-time view of internal organs, without exposing the patient to ionizing radiation, and is especially useful in obstetrics, cardiology, abdominal imaging, and emergency medicine. Ultrasound images often have high speckle noise and low image quality, and they are also operator dependent, making automated interpretation more difficult compared to other imaging modalities. AI methods have demonstrated potential in enhancing image quality, anatomical segmentation, lesion detection, and disease diagnosis, aiding clinicians in achieving precise and consistent evaluations (Rajpurkar et al., 2022; Litjens et al., 2017).

The past few years have seen digital pathology undergo a dramatic transformation since the introduction of whole slide imaging, which allows for digitization of microscopic tissue samples for computational analysis. Histopathological images are characterized by extremely high spatial resolution and therefore have a high potential for tumor detection, grading and prognosis. The large size of whole slide images, differences in tissue preparation and staining methods, and the small number of annotated images pose serious computational challenges, though. However, the development of deep learning, especially transformer-based models and foundation models, has significantly enhanced the

representation of tissues and features, paving the way for improved accuracy in the pathology workflow (Chen et al., 2026; Zhang et al., 2023).

One of the first clinical success stories of AI diagnostic assistance is through ophthalmic imaging, such as retinal fundus photography and Optical Coherence Tomography (OCT). These modalities offer in-depth imaging of the retina that allows early identification of diabetic retinopathy, glaucoma, age-related macular degeneration, and other diseases that could threaten vision. There has been some standardization of retinal data which has led to the creation of very accurate AI systems that can be used to support large scale screening programmes and enhance access to ophthalmic care (Rajpurkar et al., 2022; Litjens et al., 2017).

Endoscopic imaging can provide direct visualization of the gastrointestinal tract and is an important tool for the diagnosis of polyps, ulcers, inflammatory disorders and gastrointestinal cancers. The use of AI in endoscopy has shown significant promise in improving the accuracy of polyp detection, characterization, and real-time guidance during endoscopy. However, the challenges of continuous video, different lighting conditions, motion blur, and anatomical complexity compound the need for AI models that can handle extensive feature learning and real-time inference (Rajpurkar et al., 2022; Chen et al., 2026).

These imaging modalities have significantly different acquisition methods, imaging features and clinical utility, but present similar challenges that impact on the development of AI. Deep learning models may not be generalized due to the variability in image quality, scanner specifications, patient population, disease prevalence, annotation quality, or imaging protocols at the institution. Such distinctions have been driving the evolution of ever more advanced AI solutions, which can learn robust representations, transferable across modalities, and able to maintain reliability across a variety of clinical settings. To address these challenges and enhance the accuracy of medical image analysis using AI, deep learning architecture has emerged as a pivotal technology. Deep learning architecture has thus emerged as a key technology to overcome these limitations and drive the performance of AI-assisted medical image analysis. The following section therefore considers how current advances in medical imaging have been achieved by the development of deep learning techniques.

2.2 Evolution of AI in Medical Image Analysis

The use of artificial intelligence (AI) in medical image analysis has undergone significant changes in the last decades, from traditional image processing methods to advanced deep learning and foundation models that can learn complex visual representations from vast medical databases. This is due to the growing complexity of medical imaging data, the rising demand for automated clinical decision support, and the enhancement of computational capabilities and availability of data. Every phase of development has tackled key shortcomings of the previous generation and added new capabilities for the use of AI in illness detection, diagnosis, segmentation and prediction of prognosis (Litjens et al., 2017; Chen et al., 2026).

The initial computer-aided diagnosis (CAD) systems were mainly based on conventional image processing and machine learning methods. For these approaches, domain experts had to manually define features that describe the texture, intensity and shape of images as well as edges prior to classifying patterns of pathological images. These methods showed some promise in helping the clinician, but they needed extensive feature engineering to be successful and were hard to transfer to other imaging modalities, diseases, and clinical settings. The diagnostic performance was frequently sub-optimal, and they were not widely used in clinical practice due to differences in image quality, acquisition protocols and patient characteristics (Litjens et al., 2017).

Medical image analysis is a major field of application of deep learning, especially Convolutional Neural Networks (CNNs) which now replace handcrafted feature extraction with automated hierarchical feature learning. CNNs were able to learn from raw images, with their performance in image classification, object detection, segmentation and disease diagnosis greatly enhanced. They were found to be very competitive in many imaging modalities and are widely used for tasks like tumor detection, organ segmentation, identification of pulmonary diseases, screening of retinal diseases and dermatological diagnosis (Litjens et al., 2017; Estava et al., 2017). With the advent of specialized architectures like U-Net, biomedical image segmentation gained another significant step, integrating precise localization with the extraction of contextual features, setting the standard for contemporary segmentation networks (Ronneberger et al., 2025).

Although successful, CNN-based models have inherent limitations in modelling long-range spatial relations due to their local receptive fields. This challenge inspired the use of transformer architecture, which was initially used in natural language processing, in computer vision. Self-attention mechanisms were introduced by Vision Transformers (ViTs), which allows models to spatially attend to the global context across the whole image in challenging medical imaging applications such as those using large anatomical structures and subtle pathological patterns (Dosovitskiy et al., 2020). Further, transformer-based medical architectures, like UNETR, further merged transformer encoders into the

segmentation network to boost 3D medical image segmentation while maintaining local and global contextual information (Hatamizadeh et al., 2022). Self-supervised learning (SSP) techniques have also improved the performance of transformers, enabling them to learn informative representations from vast archives of unlabeled medical images without relying on the costly expert annotations (Tang et al., 2022).

The next phase of this evolution is the advent of foundation models which are pretrained on a vast amount of data and then fine-tuned for a wide range of medical imaging applications that are downstream of the foundation model. In contrast to task specific models, foundation models learn generalized visual representations that can be transferred across many different imaging modalities and clinical applications by requiring little to no further training. Segment Anything (SAM) was introduced for prompt-based segmentation that could be generalized to a wide variety of visual domains, while its medical adaptation, MedSAM, showed better segmentation performance on various medical images (Kirillov et al., 2023; Ma et al., 2024). Likewise, multimodal foundation models such as BiomedCLIP and LLaVA-Med have been developed to integrate visual and textual data for understanding medical images, clinical reasoning, and medical question-answering, reflecting the progress of the fusion of computer vision and large language models in the medical field (Zhang et al., 2023; Li et al., 2023).

These technological advancements have significantly enhanced the diagnostic capabilities but have also added more complications to the lives of the models, such as its issues of interpretability, robustness, fairness, privacy and clinical validation. This has led to a growing focus on the creation of trustworthy AI systems that are not only accurate and reliable but also transparent and safe, and deployed ethically. Recent research has increasingly turned to the development of trustworthy AI systems that are accurate and reliable, as well as safe and transparent, especially regarding their ethical deployment. The paradigm change is driven by the understanding that the implementation of algorithms alone is not sufficient for successful clinical adoption; it is equally important for clinicians to comprehend, trust, and integrate AI systems into their daily work to ensure their safe use in healthcare (Amann et al., 2020; WHO, 2024; Chen et al., 2026).

2.3 AI-Driven Clinical Decision Support Systems

Clinical Decision Support Systems (CDSS) are computer-based systems that provide actionable information, evidence-based recommendations, and analytical support to the healthcare professional in the context of clinical decision making and problem solving for the specific patient. Artificial intelligence has greatly enhanced the capabilities of these systems, allowing for automated interpretation of medical images, prediction of disease outcomes, prioritization of urgent cases, and assistance in treatment planning. AI-powered CDSS are not intended to fully replace clinicians but rather act as a smart assistant to assist in clinical judgment by improving the consistency of the diagnosis, reducing interpretation time and aiding in the informed decision-making process (Rajpurkar et al., 2022; Hartsock et al., 2026).

Medical image analysis is one of the most successful applications of AI based clinical decision support systems, as imaging is a major part of the process of diagnosing many diseases. In the field of clinical applications, deep learning models are increasingly being integrated into clinical workflows for the detection of abnormalities, the quantification of disease severity, the segmentation of anatomical structures and prediction of diagnosis from MRI, CT, X-ray, ultrasound, histopathology, retinal images, and endoscopic images. These systems can help doctors to process large volumes of medical images quickly, free up time for manual tasks, and identify abnormal areas that might need more in-depth study (Hartsock et al., 2026; Litjens et al., 2017).

Clinical decision support systems with artificial intelligence have been seen to have significant advantages in several medical specialties. AI is used in radiology to aid in identifying fractures, pulmonary disease, intracranial hemorrhage, and malignant lesions, as well as in prioritizing images and optimizing workflow. Pathology: AI can be used in automated analysis of whole slide images to aid in detecting and grading tumors. AI applications have proven beneficial in ophthalmology with diabetic retinopathy and glaucoma screening at expert level, and in dermatology with the deep learning model for skin lesion classification. AI has proven effective in ophthalmology, for screening diabetic retinopathy, and in dermatology, for skin lesion classification at expert level. The applications mentioned above show the potential for AI to augment the skills of specialists, enhance diagnostic efficiency, improve reproducibility, and provide more readily accessible access to high-quality healthcare services (Hartsock et al., 2026; Estava et al., 2017; Topol, 2019).

While these benefits are significant, integrating AI into clinical decision-making requires more than just a high hit rate. Healthcare professionals need to appreciate the rationale behind the recommendations made by AI models and be aware of scenarios where models might fail, as well as having adequate oversight into the decision-making process. Explainability, robustness, algorithmic bias, patient privacy, and model generalizability continue to be important challenges to clinical adoption. In addition, AI systems need to be valid across various healthcare facilities, patient populations, and imaging equipment, and need to meet regulatory standards for safety, accountability, and ethical application (Amann et al., 2020; Ueda et al., 2024; Joshi et al., 2024; WHO, 2023).

In response to these challenges, international organizations and regulatory bodies have established guidelines and frameworks for the responsible development and utilization of AI-driven clinical decision support tools. To address these challenges, international organizations and regulatory bodies have created guidelines and frameworks for the responsible development and deployment of AI-driven clinical decision support systems. Human oversight, transparency, accountability, and patient safety are a key part of the AI lifecycle, and the FDA has provided FDA pathways for AI/ML-powered medical devices. Likewise, the NIST AI Risk Management Framework and the European Union AI Act offer extensive guidance on how to identify, assess, and address the risks of AI systems used in sensitive areas like health care (WHO, 2024; Joshi et al., 2024; NIST, 2023; Act, 2024). These governance efforts emphasize the need for the creation of accurate and reliable AI systems that are clinically safe, trustworthy, and transparent.

AI-based clinical decision support is a rapidly changing field, and the fact that technology alone is not enough to make a successful implementation in healthcare. With the growing sophistication of AI models, driven by transformer architectures, foundation models, and multimodal learning, efforts need to be made to ensure explainability, robustness, fairness, privacy, and clinical applicability of these systems. In the next section we will be discussing higher-level deep learning models and methods, and these principles are relevant to trustworthy AI in general.

2.4 Deep Learning Techniques for Medical Image Analysis

The automatic learning of complex hierarchical representations using imaging data directly is driving the use of deep learning as the standard method for medical image analysis. Deep learning architectures have the ability to learn discriminative features across several layers of nonlinear transformations, and they often outperform traditional machine learning approaches in image classification, image segmentation, object detection, image registration, image reconstruction, and disease prediction for various medical imaging modalities, such as computed tomography, magnetic resonance imaging, and ultrasound (Litjens et al., 2017; Rajpurkar et al., 2022). These have advanced computer-aided diagnosis and laid the foundation of deep learning as the backbone of today's clinical decision support systems.

The development of deep learning has yielded more complex architectures to tackle specific tasks in medical image analysis. Then, Convolutional Neural Networks (CNNs) achieved successful feature learning for medical imaging in a complete end-to-end fashion, and a new type of network called an encoder-decoder network (e.g., U-Net) revolutionized biomedical image segmentation by enabling precise localization and contextual feature extraction. Thereafter, transformer-based models have enhanced the ability to model long-range spatial dependency, and foundation models have been pretrained on large-scale datasets and shown to have great cross-task generalization ability in imaging. New challenges regarding computational complexity, interpretability, robustness, and clinical deployment have been introduced by each generation of models, which provide unique strengths. These key deep learning methods are explored in detail, their clinical uses are discussed, the pros and cons are summarized, and their relevance to trustworthy medical image analysis AI systems are reviewed in the following subsections.

3 CONVOLUTIONAL NEURAL NETWORKS (CNNs)

The dawn of medical image analysis using CNNs marked a significant shift in how medical images were analyzed, as these networks automate feature extraction from medical images instead of relying on handcrafted features in traditional machine learning. Hierarchical feature learning allows CNNs to be very successful in image classification, detection and diagnosis in a variety of imaging modalities, including simple edges and complex anatomical structures (Litjens et al., 2017).

The introduction of various architectural breakthroughs like AlexNet, VGG, ResNet, DenseNet, and EfficientNet greatly enhanced feature representation, depth, and computational efficiency, making CNNs the standard in MRI, CT, X-ray, ultrasound, retinal imaging, histopathology, and dermatology (Rajpurkar et al., 2022; Hartsock et al., 2026). They showed clinical promise by being able to be classified as skin cancers by a dermatologist, and they were able to reach the level of a dermatologist in conditions that are very well controlled (Esteva et al., 2017).

While these achievements are impressive, CNNs are mainly useful for modelling local spatial structures, which prevents them from modelling long-range contextual structures in complex medical images. They also suffer from a need for large amounts of annotated data, lack of interpretability, and being sensitive to adversarial examples and domain shifts, which makes them less reliable in real-world, clinical applications (Apostolidis & Papakostas, 2021; Amann et al., 2020; Zhou et al., 2022). These restrictions drove the evolution of segmentation networks, transformer-based systems, and foundation models that better leverage context to learn, generalize, and apply to clinical problems.

3.1 U-Net and Deep Learning Segmentation Networks

Medical image segmentation is a crucial component of medical image analysis that facilitates accurate delineation of anatomical structures, lesions and pathological areas, which aid in the diagnosis, treatment planning, and monitoring of

disease. The adoption of U-Net marks a significant advancement in the ability to accurately segment relatively small annotated medical datasets despite having both high level contextual information and fine spatial details (Ronneberger et al., 2025).

U-Net has been successfully used for many medical imaging applications such as tumor segmentation, organ delineation, blood vessel extraction and lesion detection for MRI, CT, ultrasound, retinal and histopathological images (Litjens et al., 2017). The versatility and the ability to segment effectively have led to many modifications that extend its utility for better feature extraction, attention mechanisms and computation speed and adaptability to more complex clinical use (Chen et al., 2026).

U-Net inspired this, transformer-based segmentation networks like UNETR combine self-attention mechanisms with encoder-decoder designs, enabling them to grasp both local anatomical features and extended context. The combination of these models has been shown to achieve better segmentation accuracy in challenging 3D medical image segmentation tasks, especially those that involve capturing the broader context of images, which is difficult for CNN-based segmentation methods to handle (Hatamizadeh et al., 2022).

Although they are effective, segmentation networks still require good quality pixel-level annotations, which are expensive and time-consuming in clinical practice. They may also fail to perform well on images obtained from other institutions or imaging protocols and so require even more powerful and transferable segmentation methods. The constraints have led to the emergence of transformer architectures and foundation models that can acquire more complex representations and be applied to multiple medical imaging applications (Chen et al., 2026; Su et al., 2024).

3.2 Vision Transformers for medical image analysis

While CNNs have been very successful in medical image analysis, their local convolutional operations are not sufficient to capture long-range spatial relationships in complicated medical images. However, Vision Transformers (ViTs) overcome this deficit by using global dependencies through self-attention modules that process information across the entire image, resulting in more comprehensive context and representation of features (Dosovitskiy et al., 2020).

Vision Transformers have been shown to work well in various applications such as image classification, segmentation, and the detection of diseases, especially in large datasets since their introduction. Transformer-based architecture has enhanced the analysis of medical images such as MRI, CT, histopathology, and retinal imaging, by capturing both local anatomical details and global contextual information. This capability is further enhanced in models like UNETR, which incorporate the transformer encoders alongside segmentation networks, leading to improved three-dimensional medical image segmentation performance (Hatamizadeh et al., 2022; Chen et al., 2026).

Self-supervised learning techniques have recently been developed that have expanded the capabilities of transformers further by providing them with meaningful visual representations from large databases of unlabeled medical images. It minimizes the reliance on costly expert annotations and enhances the model's generalizability to a wide range of imaging applications and medical centers (Tang et al., 2022).

Despite these benefits, Vision Transformers typically need significantly more training data and processing power than CNNs. Their growing complexity in architecture also introduces challenges in terms of interpretability, computational efficiency, and clinical application, especially in healthcare settings with limited resources (Chen et al., 2026). However, their impressive learning capabilities have laid the groundwork for the future of medical AI, such as large-scale foundation models that can handle multiple imaging tasks and learn through transfer learning and multimodal reasoning.

3.3 Foundation Models and Vision-Language Models

Based on their ability to learn general purpose image representations from vast amounts of data that can be transferred to many downstream applications with limited extra training, foundation models represent the most recent step forward in medical image analysis. These architectures can support transfer learning between various imaging modalities and clinical tasks, as opposed to developing separate models for individual tasks. Foundation models have emerged as a viable option for scalable and adaptable AI-powered healthcare due to their power of generalization across datasets (Chen et al., 2026).

The most impactful developments are those of the Segment Anything Model (SAM), which enabled prompt-based image segmentation that required little task-specific training and identified objects. This medical adaptation, MedSAM, is extended to medical imaging, showing better segmentation results in MRI, CT, ultrasound and other imaging modes (Kirillov et al., 2023; Ma et al., 2024). These models are designed to minimize the amount of labeled data and enhance versatility in various clinical scenarios.

With the advent of multimodal vision-language models, the potential of AI in healthcare has been further broadened. BiomedCLIP and LLaVA-Med are two models that combine medical images with textual data to understand and reason about the images and answer related questions in a unified architecture. This multimodal learning approach can enable a more comprehensive clinical decision support system, as it integrates visual evidence with domain knowledge, thus making AI systems closer to the clinicians' clinical decision-making process in complex clinical cases (Zhang et al., 2023; Li et al., 2023; Sun et al., 2024).

While they offer significant potential, foundation models also come with new challenges. Large computation needs, reliance on huge amounts of pre-training data, lack of explainability and risks for propagating bias are issues of concern for transparency, fairness and safe clinical deployment. In addition, strict external validation and regulatory regulation are still necessary prior to their extensive adoption in routine healthcare practices (Chen et al., 2026; Chen et al., 2024). In the future, it will be essential for foundation models to be interpretable, strong, and clinically trustworthy to be widely used in medical image analysis.

3.4 Comparative Analysis of Deep Learning Techniques

Deep learning architectures have evolved to address challenges in medical image analysis, aiming to enhance feature representation, contextual understanding, and generalization. CNNs laid the groundwork for automated image interpretation, and U-Net and its successors have been developed to solve the problem of accurate medical image segmentation. In recent years, self-attention mechanisms have been introduced in Vision Transformers to better model the global context in images and recently, foundation models have further extended these capabilities by leveraging large-scale pretraining and multimodal learning. Each generation has been better at diagnosis, but no single architecture is optimal as the choice of the model depends on the imaging modality, the clinical task, the data available, the computational resources available, and deployment requirements. Table 1 summarizes the major characteristics of the deep learning techniques discussed in this section.

Table 1: Characteristics of deep learning techniques

Technique	Key Strengths	Major Limitations	Typical Applications	Trustworthiness Considerations
CNNs	Efficient feature extraction, high classification accuracy	Limited global context, dependency	Classification, detection	Explainability, robustness, domain shift
U-Net & variants	Precise image segmentation, effective with limited data	Requires detailed annotations	Organ, tumour, and lesion segmentation	Generalization, annotation quality
Vision Transformers	Global contextual learning, strong representation capability	High computational cost, large data requirements	Classification, segmentation, 3D imaging	Interpretability, computational efficiency
Foundation & Vision-Language Models	Strong transfer learning, multimodal reasoning, multi-task capability	Complex training, limited explainability, regulatory concerns	Segmentation, diagnosis, clinical reasoning	Transparency, fairness, privacy, governance

Though deep learning has developed from task-specific architectures to more general and versatile models that can be used across multiple clinical tasks. Despite the advancements in predictive performance, however, there are also challenges around explainability, robustness, fairness, privacy, and regulatory compliance. These are key challenges that need to be overcome if these high-performance AI models are to be implemented in safe and reliable clinical decision support systems. The following section therefore explores the key principles of Trustworthy Artificial Intelligence and how they contribute to responsible use of deep learning models in medical image analysis.

4 TRUSTWORTHY ARTIFICIAL INTELLIGENCE IN MEDICAL IMAGING

While deep learning has achieved significant success, the use of AI in medical image analysis is still rapidly gaining traction, but its ability to achieve high predictive accuracy is not enough for clinical use. Healthcare is a high-risk application area where errors in prediction, suggestions that may be misleading, or incorrect model behavior, can have direct effects on patients. The focus now has shifted from building more accurate models to building more trustworthy

and reliable AI systems that are explainable, robust, fair, transparent, privacy-preserving, and ethical and regulatory compliant (WHO, 2024; NIST, 2023).

The goal of trustworthy AI is to help empower clinical decision-making by allowing healthcare professionals to comprehend the model's predictions, evaluate their reliability, and recognize any constraints prior to implementing the recommendations provided by AI. The use of AI responsibly demands ongoing risk management, oversight of the human user, transparency, accountability and monitoring throughout the AI lifecycle (WHO, 2023; Joshi et al., 2024; Act, 2024).

The principles of Trustworthy AI are tightly woven together. Explainable models boost clinician confidence, robust models ensure successful performance across diverse clinical scenarios, fair models minimize demographic bias, privacy-preserving models safeguard sensitive patient information, and regulatory compliance supports safe and ethical usage. These principles form the basis for clinically reliable medical image analysis aided by AI. Each of these aspects and their importance for clinical decision support are discussed in more detail in the following subsections.

4.1 Explainability and Interpretability

Opacity of complex models is one of the biggest hurdles for clinical use of deep learning. There are many AI systems that perform well but are “black boxes” and difficult to understand how they make predictions. Explainability plays a crucial role in healthcare, where the interpretation of the results of the diagnostic process can ultimately impact patient care, clinical validation and regulatory clearance (Amann et al., 2020; Holzinger et al., 2019).

There are several explainability techniques that have been developed to increase model transparency. Grad-CAM is an effective tool for understanding the image regions that are important for the CNN-based medical image classifiers and can be used to create visual heat maps (Selvaraju et al., 2017). Explanations using model-agnostic approaches like LIME and SHAP (Ribeiro et al., 2016; Lundberg & Lee, 2017) complement those provided by the models and estimate the contribution of individual features to the model prediction, regardless of the model architecture.

These methods make the models more interpretable but may not make sense of the reason behind the functioning of AI systems. Recent research warns that the explanation methods might lead to visually convincing outputs that do not align with the way a model reaches a prediction, which could result in a false sense of confidence (Ghassemi et al., 2021). Therefore, explainability should be considered as part of the trustworthy AI, not the final answer. Future approaches need to integrate model design that is interpretable to the modeler and communicate to the clinician, and rigorous validation to make sure that the explanation is meaningful and reliable.

4.2 Robustness and Generalization

To ensure clinical reliability, AI systems need to demonstrate consistent performance across different settings, not just during development. Medical imaging introduces distribution shifts that can have a significant impact on model performance due to a variety of reasons, such as differences in imaging protocols or scanners, patient demographics, or healthcare institutions. As a result, robustness and generalization have emerged as key needs of trustworthy AI, maintaining the accuracy and reliability of models in various clinical settings (Zhou et al., 2022; Su et al., 2024).

Deep learning models are also subject to adversarial perturbations, imaging artifacts, and noise, and can make errors even when the perturbations or artifacts are not detected by clinicians. These vulnerabilities introduce patient safety issues, especially when AI tools are not properly validated and used in clinical decision-making. Resilient model training with adversarial defense strategies, domain adaptation, and external validation at multiple institutions have therefore been highlighted recently to enhance model resilience and minimize performance degradation in real-world deployment (Apostolidis & Papakostas, 2021).

The development of more robust algorithms is not enough; the entire process must be assessed with a variety of patients and imaging conditions. A good generalizable model will have better chances for successful implementation and to avoid unforeseen problems after deployment across hospitals and imaging devices. Robustness should be viewed as a necessary feature rather than a desired feature of trustworthy AI, therefore (Su et al., 2024; WHO, 2023).

4.3 Fairness and Bias Mitigation

Fairness is an important aspect of trustworthy AI, as inequitable algorithms can create inequitable health outcomes among patients in different groups. Bias may be created by an imbalance in the data sets, in disease rates, in the demographics, or in imaging equipment and clinical procedures. Failure to mitigate these biases can lead to disproportionate performance by AI systems, contributing to healthcare inequities that may be entrenched (Ueda et al., 2024).

AI models have been demonstrated to have underdiagnosis bias, especially when trained on data that is not representative of all patient groups. These biases can affect the accuracy of the diagnosis for specific groups and create lack of trust in AI support for decision-making (Seyyed-Kalantari et al., 2021). To overcome these challenges, there is a need for a variety of representative datasets, model development that is fair, continuous monitoring of model performance and clear reporting of model performance across patient populations.

Building fairness should not be considered only a technical goal but an ethical and clinical duty. To guarantee equitable model performance following deployment, it is critical that the model is constantly evaluated throughout its entire lifecycle, from development to application. Healthcare organizations can minimize algorithmic bias and enhance the reliability and inclusivity of medical imaging systems powered by AI by incorporating fairness assessments during the model development, validation, and clinical monitoring stages (WHO, 2024; NIST, 2023).

4.4 Privacy-Preserving AI

Patient privacy is paramount in medical imaging data, and safeguarding it is crucial for the trustworthiness of AI. Historically, the model training process used a centralized model training system that demanded the centralized collection of a huge volume of patients' data, which led to the possibility of data theft, unauthorized access and non-compliance with regulations. With a growing interconnectedness of healthcare organizations, new solutions have emerged to replace the collaborative development of AI through privacy-preserving learning strategies (Zhu et al., 2024).

Federated learning is one of these methods, which allows different institutions to train AI models together without sharing patient data with other institutions. Institutions do not exchange raw medical images, but rather they exchange model parameters which still maintain patient privacy but also provide a diverse set of images. This has shown to be a very promising way to enhance model generalization, enable collaboration between multiple institutions and speed up the development of clinically viable AI systems (Sheller et al., 2020; Rieke et al., 2020).

While federated learning is a solution to many privacy concerns, it's not a solution to all security concerns. The challenges of model inversion attacks, communication and data heterogeneity are still significant and need to be addressed. Therefore, privacy-preserving AI systems should also be designed with strong security protocols, ethical governance frameworks, and regulatory compliance to safeguard patient privacy throughout the AI lifecycle (Zhu et al., 2024; Utomi et al., 2024).

4.5 Reliability, Transparency, and Regulatory Compliance

While reliable computer systems are required to make consistent and reproducible predictions of clinical value, sufficient transparency should be provided for clinicians to understand the intended uses and limitations of the computer system. Reliability is not only about predictive accuracy, but also robust validation in various clinical settings, ongoing performance monitoring and human supervision during implementation. Even very accurate models can fail when transferred to use in a routine health care setting if they are not adequately protected (WHO, 2023; NIST, 2023).

Transparency is related to reliability by ensuring that the model development process and application of the model are well documented, including intended uses, training data, model performance, and limitations. This type of data helps healthcare professionals, organizations, and regulators make decisions about implementing AI and holds those who use AI systems accountable when they impact patient care. Additionally, transparent reporting ensures that clinicians have a clear understanding of the situations in which AI recommendations are accepted, considered, or rejected, enhancing their trust in the tool (WHO, 2024).

As AI systems move from research to clinical settings, regulatory oversight is gaining significantly. WHO, FDA, NIST, and the European Union have created frameworks that highlight the importance of risk management, patient safety, accountability, manual oversight, and ongoing post-deployment monitoring of AI systems in healthcare (Joshi et al., 2024; WHO, 2023; Act, 2024; NIST, 2023). These complementary findings further underscore the critical role of data protection, legal frameworks, and governance in responsible AI use within healthcare systems (Korang et al., 2024; Osifowokan et al., 2025; Utomi et al., 2024).

Reliability, transparency, and regulatory compliance work together to ensure the governance needed to shape AI innovations into safe and trustworthy clinical practice. In the era of ever more sophisticated deep learning models, these principles will continue to be crucial to guarantee that AI-driven medical image analysis has clinically relevant value and retains the public's trust.

5 APPLICATIONS IN CLINICAL DECISION SUPPORT

AI has emerged as a key part of clinical decision support, providing healthcare workers with tools to detect, diagnose, segment, prognosis and monitor diseases from medical images. The models of deep learning have shown extraordinary success in many medical specialties, learning MRI, CT, X-ray, ultrasound, histopathological, retinal and endoscopic images with increasing accuracy and efficiency. These systems aren't meant to supplant the work of clinicians, but rather to act as decision support tools that augment clinical judgment to shorten interpretation time, increase the consistency of clinical interpretation, and prioritize cases that need urgent medical attention (Rajpurkar et al., 2022; Hartsock et al., 2026).

One of the most developed uses of AI-powered medical imaging is radiology, where deep learning algorithms can be used to identify fractures, lung conditions, intracranial bleeding, cardiovascular problems, and cancers out of MRI, CT and X-ray scans. Automated image analysis allows for quick detection of suspicious findings and quantitative analysis of disease progression, streamlining workflow and lowering clinician burden. In oncology and digital pathology, AI algorithms are used to analyze high-resolution histopathological images, helping to detect segment, grade tumors, and plan treatment options. In oncology and digital pathology, AI algorithms analyze high-resolution histopathological images to aid in tumor detection, segmentation, grading, and treatment planning. The rise of transformer architectures and foundation models has further enhanced the ability to analyze complex pathology images, capturing both the local characteristics of tissues and the broader context (Hartsock et al., 2026; Chen et al., 2026).

In ophthalmology and dermatology, AI has made major strides in the automated identification of diabetic retinopathy, glaucoma, age-related macular degeneration, and skin cancer. The capabilities of deep learning have been showcased in landmark studies of its ability to achieve dermatologist-level performance in skin lesion classification, which holds potential to improve early disease detection and help extend access to specialist care (Estava et al., 2017). More recently, foundation models and vision-language architectures, like BiomedCLIP (Zhang et al. 2023) and LLaVA-Med (Li et al., 2023), have extended the capabilities of AI beyond image interpretation by combining visual information with medical knowledge, so that multimodal clinical reasoning and medical question answering are possible.

Though these strides have been made, more than an accurate diagnosis is required to have successful clinical implementation. The recommendations from AI systems need to be reliable, explainable and clinically relevant, and should be interpretable and accepted by physicians before entering clinical care. Therefore, the key tenets of AI, such as explainability, robustness, fairness, protection of privacy and compliance to regulatory frameworks continue to be core to AI clinical decision support's success in enhancing patient outcomes and retaining clinician trust and patient safety (Amann et al., 2020; WHO, 2023).

6 CHALLENGES AND LIMITATIONS

Although deep learning has made great strides in medical image analysis, there are still several challenges that may impede its clinical use. While most AI models are trained and tested with meticulously selected data samples in research settings, healthcare contexts are significantly more complex and varied in the real world. Model performance is often impacted by distribution shifts arising from variations in imaging protocols, scanner manufacturers, patient demographics, disease prevalence, institutional practices, and any other factors that may differ from the training setting. Additionally, the lack of access to large, well-annotated medical imaging databases and privacy concerns involving sharing data persist and further limit model development and generalization across healthcare institutions (Litjens et al., 2017; Zhou et al., 2022; Su et al., 2024).

The other big obstacle is building clinician trust. While techniques for explaining model predictions like Grad-CAM, SHAP, and LIME offer insights that are useful, they are not comprehensive enough to capture the reasoning processes of complex deep learning models. As a result, there is a need for a more comprehensive and consistent explanation, as well as an explanation that is easier to grasp, that will make it easier for clinicians to trust and accept AI recommendations. Therefore, when the explanation is incomplete, inconsistent, or difficult to understand, clinicians may be reluctant to trust and accept AI recommendations. Reliability is also impacted by adversarial attacks, noisy images, and other variations that can cause a lot of degradation in the system's performance, and therefore may affect patient safety, as well as the robustness of the AI supported decision making (Selvaraju et al., 2017; Ghassemi et al., 2021; Apostolidis & Papakostas, 2021).

Implementation has ethical and operational challenges. An AI system that has been trained on an unrepresentative sample can have a demographic bias, which can lead to different diagnostic performance in different patient groups. Concurrently, the application of medical imaging data in sensitive contexts highlights significant privacy and security issues that demand strong measures across the entire AI lifecycle. In spite of privacy-preserving methods like Federated Learning, which enable collaborative model building without the sharing of patient information, however, there is a

need for further protection to address security concerns and regulatory compliance in healthcare (Ueda et al., 2024; Seyyed-Kalantari et al., 2021; Sheller et al., 2020; Zhu et al., 2024).

In addition to technical issues, there are also significant challenges related to the integration of AI into clinical care, such as the need for substantial computing resources, interoperability with existing healthcare systems, ongoing monitoring of models, and training of clinicians. AI's advancement has also surpassed the capabilities of current regulatory frameworks, which leaves questions about standards for validation, accountability, and ongoing supervision. The WHO, FDA, NIST and the European Union have set up important governance frameworks, but ongoing cooperation between researchers, clinicians, healthcare institutions, and policymakers is crucial to build a common framework for evaluating systems and ensure that AI-assisted medical imaging systems are deployed safely, ethically, and sustainably (WHO, 2023; Joshi et al., 2024; NIST, 2023; Act, 2024).

Building trustworthy AI requires the cooperation of researchers, clinicians, regulators, and healthcare organizations to develop standard evaluation procedures, ongoing monitoring systems, and governance frameworks that will guarantee the safety, effectiveness, and accountability of AI systems as they are put into practice. Solving these issues will be essential to making the progress of deep learning available in sustainable and reliable clinical decision support.

7 EMERGING TRENDS AND FUTURE RESEARCH DIRECTIONS

The field of medical image analysis is ripe for the application of artificial intelligence, which is advancing with each passing day thanks to the progress of deep learning architectures, computational power, and the abundance of massive medical data. Recent works have moved away from task-specific models to more general and flexible AI systems that can be applied to multiple clinical applications. Concurrently, there is an increasing focus on making these systems explainable, robust, fair, privacy preserving and following ethical and regulatory frameworks. Moving forward, ensuring predictive success will be dependent on the development of AI systems that can be safely and reliably introduced into standard clinical practice will also be necessary.

The rise of foundation models and multimodal AI is one of the most important innovations. Large-scale pretraining models like SAM, MedSAM, BiomedCLIP, and LLaVA-Med illustrate the promise of transferable feature representations and multimodal learning in the realm of medical imaging for a diverse array of tasks (Chen et al., 2026; Ma et al., 2024; Zhang et al., 2023; Li et al., 2023). These models can combine medical images with textual information and extend beyond image interpretation to clinical reasoning and clinical decision support. They are anticipated to be more efficient, interpretable, clinically valid, and ultimately have the capability to integrate more patient information, including EHRs, lab results, and clinical notes.

In addition to these architectural developments, the focus has shifted to building inherently trustworthy AI systems. Future models should offer explanations that are meaningful to the clinic, consistent performance across health care environments, and be able to reduce demographic bias via representative datasets and on-going fairness checks. There will also be significant research efforts to further enhance robustness to adversarial attacks, imaging artefacts, and distribution shifts, which directly affect the reliability of AI-assisted clinical decisions (Amann et al., 2020; Apostolidis & Papakostas, 2021; Ueda et al., 2024). Future research should focus on solving the problem of explainability, robustness and fairness not separately, but rather along the whole AI development lifecycle.

Collaborative learning and privacy-preserving will also be more prevalent to support trustworthy AI. Federated learning has been shown to be a method for creating high-performing AI models across several healthcare institutions without the sharing of patient-specific, sensitive information, thus enabling the model to be more generalized while maintaining patient privacy (Sheller et al., 2020; Rieke et al., 2020). As privacy-preserving learning continues to evolve and multi-institutional datasets grow, and evaluation protocols become standardized, the scalability and reliability of AI systems in healthcare settings are likely to grow.

In conclusion, a joint effort between researchers, doctors, health-care institutions and authorities will be essential for the future success of AI in medical image analysis. The WHO, FDA, NIST, and European Union have developed frameworks that offer valuable guidance for responsible development of AI across the AI lifecycle, starting with transparency and accountability, moving to human oversight, risk management, and ending with continuous monitoring (WHO, 2023; Joshi et al., 2024; NIST, 2023; Act, 2024). Additional work is thus needed to expand the capabilities of deep learning models, while also ensuring that these technologies are clinically validated, ethically regulated, and can be used to enable safe, equitable, and trustworthy clinical decision support. By integrating advances in foundation models, multimodal learning, explainability, robustness, fairness, privacy preservation, and regulatory compliance, the next generation of AI systems will be better positioned to improve diagnostic accuracy, strengthen clinician confidence, and enhance patient outcomes.

8 CONCLUSION

Through increasingly complex deep learning models, artificial intelligence has made great strides in medical image analysis, enhancing disease detection, segmentation, diagnosis, and clinical decision support. While models like CNNs, Vision Transformers and foundation models have demonstrated impressive performance, their clinical success hinges not just on predictive accuracy but on other factors. This review highlights the capabilities of these architectures, while also revealing their limitations. To establish trustworthy AI systems that can be confidently incorporated into healthcare practice, it is crucial to create systems that are explainable, robust, fair, preserve privacy, reliable, and compliant with regulations. Despite the problems with data quality, data generalization, bias and governance, the progress in multimodal learning, foundation models and collaborative AI hold promises of more robust and patient-centered solutions. Moving forward, it is important for further studies to focus on fostering a balance between technological advancements and ethical and regulatory frameworks to ensure that the use of AI systems for medical image analysis is safe, transparent, and trustworthy, ultimately contributing to better patient outcomes.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Disclosure of conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

- [1] Act, E. A. I. (2024). The eu artificial intelligence act. European Union.
- [2] Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & Precise4Q Consortium. (2020). *Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making*, 20(1), 310.
- [3] Apostolidis, K. D., & Papakostas, G. A. (2021). A survey on adversarial deep learning robustness in medical image analysis. *Electronics*, 10(17), 2132.
- [4] Chen, Q., Liu, A., Zhang, J., Yang, C., & Zhang, Y. (2026). Foundation models in medical imaging: A review. *EngMedicine*, 3(2), 100123.
- [5] Chen, X., Xiang, J., Lu, S., Liu, Y., He, M., & Shi, D. (2024). Evaluating large language models in medical applications: a survey. *arXiv preprint arXiv:2405.07468*.
- [6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- [7] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639), 115-118.
- [8] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The lancet digital health*, 3(11), e745-e750.
- [9] Hartsock, I., Koutsoubis, N., Ahmed, S., Parker, N., Schabath, M. B., Araujo, C., ... & Rasool, G. (2026). Clinical AI in Radiology: Foundations, Trends, Applications, and Emerging Directions. *Cancers*, 18(6), 942.
- [10] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., ... & Xu, D. (2022). Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 574-584).
- [11] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 9(4), e1312.
- [12] Joshi, G., Jain, A., Araveeti, S. R., Adhikari, S., Garg, H., & Bhandari, M. (2024). FDA-approved artificial intelligence and machine learning (AI/ML)-enabled medical devices: an updated landscape. *Electronics*, 13(3), 498.
- [13] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015-4026).
- [14] Korang, A., Osifowokan, A. S., Seabac, V. S., & Donkor, A. A. (2024). An Assessment of AI Regulatory and Compliance Issues in the US Healthcare Sector.
- [15] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., & Gao, J. (2023). LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. 10.48550/arXiv.2306.00890.

- [16] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42, 60-88.
- [17] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- [18] Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment anything in medical images. *Nature communications*, 15(1), 654.
- [19] NIST. (2023). Artificial intelligence risk management framework (AI RMF 1.0). URL: <https://nvlpubs.nist.gov/nistpubs/ai/nist.ai>, 100-1.
- [20] Osifowokan, A. S., Agbadamasi, T. O., Adukpo, T. K., & Mensah, N. (2025). Regulatory and legal challenges of artificial intelligence in the US healthcare system: Liability, compliance, and patient safety. *World Journal of Advanced Research and Reviews*, 25(3), 949-955.
- [21] Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature medicine*, 28(1), 31-38.
- [22] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.
- [23] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ digital medicine*, 3(1), 119.
- [24] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer international publishing.
- [25] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).
- [26] Seyyed-Kalantari, L., Zhang, H., McDermott, M. B., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12), 2176-2182.
- [27] Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., & Bakas, S. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep*. 2020; 10 (1): 12598. Raven, W., Nissen, SK, Morsink, ME, Gaakeer, MI, Brabrand, M,... & de Groot, B, 223-230.
- [28] Su, Z., Guo, J., Yang, X., Wang, Q., Coenen, F., Hussain, A., & Huang, K. (2024). Navigating distribution shifts in medical image analysis: A survey. *arXiv preprint arXiv:2411.05824*.
- [29] Sun, G., Qin, C., Fu, H., Wang, L., & Tao, Z. (2024, November). Self-training large language and vision assistant for medical question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 20052-20060).
- [30] Tang, Y., Yang, D., Li, W., Roth, H. R., Landman, B., Xu, D., ... & Hatamizadeh, A. (2022). Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20730-20740).
- [31] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1), 44-56.
- [32] Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., ... & Naganawa, S. (2024). Fairness of artificial intelligence in healthcare: review and recommendations. *Japanese journal of radiology*, 42(1), 3-15.
- [33] Utomi, E., Osifowokan, A. S., Donkor, A. A., & Yowetu, I. A. (2024). Evaluating the impact of data protection compliance on AI development and deployment in the US health sector. *World Journal of Advanced Research and Reviews*, 24(2), 1100-1110.
- [34] World Health Organization. (2023). *Regulatory considerations on artificial intelligence for health*. World Health Organization.
- [35] World Health Organization. (2024). *Ethics and governance of artificial intelligence for health: large multi-modal models*. WHO guidance. World Health Organization.
- [36] Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., ... & Poon, H. (2023). Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

- [37] Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2022). Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4), 4396-4415.
- [38] Zhu, Y., Yin, X., Wee-Chung Liew, A., & Tian, H. (2024, December). Privacy-preserving in medical image analysis: A review of methods and applications. In *International Conference on Parallel and Distributed Computing: Applications and Technologies* (pp. 166-178). Singapore: Springer Nature Singapore.