Analysis of AI Diagnostic Performance Discrepancies Across Medical Imaging Modalities

Anonymous Author(s)
Affiliation
Address
email

Abstract

Artificial intelligence (AI) shows immense promise in medical imaging, yet its diagnostic performance varies significantly across different modalities. This discrepancy is highlighted by the "ultrasound paradox," where AI achieves superior performance on comparatively lower-quality ultrasound images (AUROC 0.94) while struggling with high-resolution, complex modalities like MRI (reported accuracy as low as 0%). This suggests that performance is not dictated by image quality alone but by a complex interplay between the data's intrinsic properties and the structural limitations of current AI architectures. This paper provides a deep-dive analysis of this performance gap by systematically reviewing literature on static, high-contrast (CT, MRI) and dynamic, low-contrast (X-ray, ultrasound) modalities. We investigate the root causes, attributing them to a mismatch between the information type provided by a modality (e.g., spatio-temporal data in ultrasound) and the architectural constraints of dominant AI models like Convolutional Neural Networks (CNNs), such as their limited receptive fields and difficulty in processing temporal features. As a practical solution, we propose a multi-stage "hybrid diagnostic workflow" that strategically combines high-sensitivity AI for initial screening (using X-ray/ultrasound) with high-specificity AI for confirmation (using CT/MRI). This approach aims to optimize overall diagnostic accuracy and clinical efficiency. We conclude that the future of medical AI lies not in a single, universal model but in an integrated, collaborative ecosystem that leverages the unique strengths of different modalities and AI architectures to create robust, clinically-relevant solutions.

1 Introduction

Artificial Intelligence (AI) is driving a revolutionary shift in medical imaging, significantly contributing to enhanced diagnostic accuracy and improved clinical workflows. Deep learning algorithms, in particular, demonstrate the ability to recognize complex patterns from large-scale datasets, achieving expert-level diagnostic performance in several domains. A framework developed at UCLA has even shown that deep learning AI can rapidly achieve clinician-level accuracy in complex medical image analysis.

The rapid advancement and practical application of medical imaging AI are evidenced by the fact that approximately 76% of the over 1,000 AI-based medical devices approved by the U.S. FDA are concentrated in radiology. For instance, large-scale studies have shown that AI assistance in breast cancer screening can increase cancer detection rates by 20-30%. In prostate cancer diagnosis, AI has demonstrated the ability to reduce the rate of missed clinically significant lesions from 8% by radiologists to just 1%. These examples underscore AI's contribution to improving diagnostic sensitivity and reading efficiency in real-world clinical settings. While AI has long demonstrated superhuman capabilities in analyzing structured numerical data, such as blood test results, its application to the unstructured and complex domain of medical imaging reveals a far more nuanced and paradoxical landscape of performance.

However, a notable issue has emerged: the performance of medical AI varies significantly depending on the imaging modality. A systematic review revealed that while ultrasound-based AI models achieved a very high mean Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.94 (95% CI 0.88–1.00), CT and MRI-based models lagged behind at approximately 0.82 (CT: 95% CI 0.78–0.86; MRI: 0.71–0.93). More strikingly, a recent evaluation of the latest ChatGPT-4 vision model reported diagnostic accuracies of around 30% for X-ray images and 40% for CT, but 0% for MRI. This **"ultrasound paradox"**—where the highest performance is observed in a modality with relatively lower image quality—provides compelling evidence that AI performance cannot be predicted by physical image quality alone. It raises a fundamental question about what kind of information AI models learn most effectively and suggests that the performance gap stems not only from the intrinsic properties of the images but also from the structural limitations of current AI architectures.

This study aims to systematically analyze the phenomenon of AI performance discrepancy across imaging modalities, identify its underlying causes, and propose practical solutions. Focusing on the performance differences between static/high-contrast (CT, MRI) and dynamic/low-contrast (X-ray, ultrasound) imaging, we explore the limitations of current AI model architectures and the potential of a hybrid approach to overcome them. Through this analysis, we seek to provide insights that go beyond technical evaluation to inform the future direction of medical AI development and its clinical application strategies.

2 AI Performance in Static/High-Contrast Imaging (CT, MRI)

2.1 AI Performance in CT Imaging

CT imaging provides favorable conditions for AI model training with its high spatial resolution and excellent tissue contrast. Deep Learning Reconstruction (DLR) techniques have demonstrated superior noise suppression and artifact reduction compared to traditional iterative reconstruction methods, enhancing image quality while reducing radiation exposure [1, 2]

For example, GE Healthcare's 'TrueFidelity' DLR system reconstructs high-quality images with over 50% less radiation, proving effective in detecting liver lesions as small as 0.5 cm. AI's role in lung cancer screening is also noteworthy [3, 4]. Recent studies show that AI systems can automatically track changes in pulmonary nodules across serial CT scans, aiding in the early detection of potentially malignant nodules and assisting clinicians in diagnosis and treatment planning [5, 6]. From an architectural perspective, the 3D volumetric data from CT is advantageous for CNNs to extract hierarchical features layer by layer [7].

However, CNNs' limited local receptive fields make it difficult to capture long-range dependencies, posing a challenge for understanding complex global anatomical relationships [3, 8, 9]. This suggests that Transformer-based models, with their ability to capture global context, could serve as a complementary solution. Indeed, in brain tumor MRI analysis, Vision Transformer (ViT) models have outperformed CNN-based models with over 98% accuracy, highlighting the importance of global information in precision diagnostics [10, 11].

2.2 AI Performance and Limitations in MRI Imaging

MRI is an essential modality for the precise diagnosis of conditions like tumors and brain diseases, thanks to its excellent soft-tissue contrast and diverse imaging sequences [12–14]. In specific, well-defined tasks, AI has shown outstanding performance [15, 16]. For instance, a ViT-based model achieved 98.5% accuracy in classifying brain tumors from MRI scans when provided with sufficient data and optimization [17, 18]. Furthermore, AI technology has been developed to reduce the use of gadolinium-based contrast agents by 80-90% while maintaining diagnostic quality, demonstrating the potential to synthesize high-quality images from low-dose contrast scans [19, 20]. This approach is significant for improving patient safety and cost-effectiveness.

Nevertheless, the complex, multi-dimensional data structure of MRI remains a challenge for AI models [21, 22]. The reported 0% diagnostic accuracy of ChatGPT-4 on MRI images underscores the failure of current general-purpose AI models to comprehend MRI's complexity [23, 24]. MRI data, which includes multiple sequences and 3D spatial information, presents a multi-dimensional problem that is difficult for traditional 2D-centric CNNs to fully capture [25]. This limitation is tied to the architectural constraints of current AI; while CNNs excel at local pattern recognition, they are weak in understanding global correlations and integrating temporal/sequential information [26], which limits their utility in multi-sequence MRI interpretation. Consequently, architectures like Transformers [27], 3D-CNNs [28], or their hybrid models are being proposed as more suitable for MRI analysis [28].

2.3 AI Performance Factors in Static/High-Contrast Imaging

The generally stable performance of AI in static/high-contrast imaging like CT and MRI can be attributed to several factors:

Structural Consistency: Human anatomical structures appear in relatively predictable and consistent forms in CT and MRI, creating feature maps that are easy for CNNs to learn.

High Signal-to-Noise Ratio (SNR): Low noise and clear contrast between tissues make it easier for AI models to distinguish features, enhancing sensitivity even for small lesions.

Standardized Acquisition Protocols: The relatively standardized and repeatable examination protocols for CT and MRI ensure consistency in training data, which improves the generalizability of the learned patterns.

Utilization of 3D Spatial Information: CT, in particular, provides 3D volumetric data, allowing models like 3D-CNNs to leverage spatial context between adjacent slices to improve diagnostic accuracy.

Thanks to these advantages, the average AUROC for CT-based AI models is reported to be around 0.82 [29], with performance comparable to specialists in tasks like tumor detection and organ segmentation [30]. While MRI performance varies by task, AI has shown expert-level results in fields like neuroimaging [31], though generalizability remains an area for improvement due to the aforementioned structural complexity [32].

3 AI Performance in Dynamic/Low-Contrast Imaging (X-ray, Ultrasound)

3.1 AI Performance and Limitations in X-ray Imaging

X-ray is the most fundamental and widely used medical imaging modality, serving as a primary examination tool in various fields. Commercial AI-assisted X-ray reading systems are already in use [33], with one independent evaluation of the Rayvolve system reporting a sensitivity of 96.4% and a specificity of 84.4% [34]. This tendency for high sensitivity coupled with somewhat lower specificity is a typical characteristic of X-ray AI [35]. A large multi-center study showed that AI assistance improved the AUC for chest X-ray interpretation by approximately 16% (from 0.759 to 0.88) and reduced reading times.

Key technical challenges for AI in X-ray imaging include:

Overlapping Structures: As a 2D projection of 3D information, X-rays suffer from information loss due to overlapping anatomical structures. This can confuse models like CNNs that extract features from local patches and lack global context [36].

Low Soft-Tissue Contrast: The low contrast of soft-tissue lesions makes it difficult for models to distinguish the boundaries and shapes of subtle abnormalities [37].

Variability in Conditions: X-ray acquisition is subject to high variability from patient positioning, exposure settings, and equipment differences, which can degrade the generalization performance of trained AI models [38].

Limitations of Local Processing: Traditional CNNs process images with local filters, making it difficult to capture widespread abnormalities or relationships between distant regions [39]. To address this, research is ongoing into Transformer-based global attention models or adding attention mechanisms to CNNs [40].

3.2 Superior AI Performance in Ultrasound Imaging

Surprisingly, AI performance in ultrasound imaging has been reported to surpass that of other modalities. The systematic review previously mentioned found that the average AUROC of 0.94 for ultrasound-based AI was significantly higher than the 0.82 for CT/MRI [41]. This suggests that the real-time nature and diverse information in ultrasound images work to AI's advantage [40]. In breast cancer diagnosis, for example, a deep learning model named DeepBreastCancerNet achieved a remarkable classification accuracy of 99.35% using ultrasound images [42].

Success factors for ultrasound AI include:

Utilization of Real-Time Dynamic Information: Ultrasound videos capture temporal changes in organ movement, lesion morphology, and blood flow signals, providing additional information not present in static images.

Compensating for Operator Dependency: AI can reduce inter-operator variability by interpreting images based on a consistent, learned standard, thereby raising the overall quality of diagnoses, especially for less experienced practitioners.

Immediate Feedback and Interaction: Real-time AI integration can provide immediate alerts for abnormalities during an examination, guiding the operator to perform additional scans or adjust angles.

Common technical challenges across dynamic/low-contrast imaging also exist:

Difficulty in Learning Spatio-Temporal Features: Traditional 2D CNNs are ill-equipped to handle the temporal dimension of dynamic videos [43, 44]. Hybrid models like CNN-LSTM are being introduced to address this. For instance, a CNN-LSTM model achieved 97.33% accuracy in predicting bone fracture healing from a series of X-rays, significantly outperforming a pure CNN [45, 46].

Noise and Artifacts: Ultrasound's speckle noise and X-ray's scatter and motion blur can degrade AI performance. Pre-processing techniques or noise-robust model architectures are essential [47, 48].

Lack of Standardization: The wide variety of equipment, settings, and protocols for ultrasound and X-ray makes it difficult for an AI model optimized in one institution to perform well in another [38]. Domain adaptation and federated learning are being explored to overcome this [49].

In summary, AI performance in X-ray and ultrasound is determined by a combination of the physical limitations of the input data and the structural constraints of current models. The exceptional performance in ultrasound paradoxically highlights these constraints, revealing the potential of AI to leverage temporal and multi-dimensional data when properly equipped.

4 A Deeper Look into the Causes of Performance Discrepancy

4.1 Hypothesis 1: The Impact of Physical Image Properties on AI Performance

The hypothesis that the physical and technical characteristics of an imaging modality directly impact AI performance is supported by numerous observations [50]. The superiority of static/high-contrast imaging, such as CT and MRI, lies in their high information richness, providing clear anatomical boundaries and relatively low noise [51, 52], which is advantageous for the local pattern learning of CNNs[53].

Conversely, the challenges in dynamic/low-contrast imaging stem from physical limitations [54]. The information loss and low soft-tissue contrast in 2D projected X-rays weaken the signal AI needs to learn from, increasing uncertainty [55]. The 30% accuracy of ChatGPT-4 on X-rays starkly illustrates the negative impact of ambiguous image features [56].

The exceptional performance of ultrasound, however, cannot be explained by traditional image quality metrics alone. Despite its noise and operator dependency, the vast number of frames and diverse acoustic information from real-time scanning appear to benefit AI [57]. This implies that even if physical image quality is lower, AI performance can be high if the quantity and type of information are rich and useful for the model.

4.2 Hypothesis 2: The Role of AI Architectural Limitations

The second hypothesis posits that performance discrepancies arise from the inherent limitations of the model architectures themselves. CNNs, the mainstream models in medical imaging AI [58, 59], have structural constraints that negatively affect their performance on certain modalities.

First is the issue of **CNN's local receptive field**. The area of an image a CNN can "see" at once is limited by its filter size and depth [4, 60], making it difficult to understand long-range relationships between distant image regions. This is a disadvantage in images covering large anatomical areas, where global context is crucial. Transformer-based models, with their self-attention mechanisms, have the potential to overcome this limitation by integrating global information [61].

Second is the **inability to learn temporal or dynamic patterns**. CNNs are designed for static 2D images and cannot capture time-varying patterns in videos like ultrasound or longitudinal image series [62]. As mentioned, the significant performance boost from using a CNN-LSTM hybrid for tracking bone healing highlights this deficit [45].

Third is the **complexity of handling multi-dimensional data**. For 3D multi-channel data like MRI, 2D CNNs struggle to extract all necessary volumetric features [63]. While 3D-CNNs exist, they are often limited by high computational costs and data scarcity [64, 65]. Recently, 3D-specific Vision Transformers and the development of large-scale "foundation models" for medical imaging are showing promise in this area.

Recent trends show a move towards **hybrid architectures** like UTNet, Swin-Unet, and ConvFormer, which combine the strengths of CNNs (local detail detection) and Transformers (global context learning) to achieve high performance more efficiently, even in low-data environments [66]

4.3 An Integrated Understanding of Performance Discrepancies

Synthesizing these two hypotheses, the performance gap across imaging modalities is best understood as an interaction between the image's characteristics and the AI model's structural properties.

Information Richness vs. Information Comprehension: CT/MRI provide physically rich information, but current models may not fully utilize it [67]. Conversely, ultrasound may have less information in terms of resolution but provides it in a form (real-time change) that models can effectively leverage [68].

Lack of Modality-Specific Architectures: Most medical AI has been developed using CNNs optimized for static 2D images. This creates a performance deficit for modalities where 3D or temporal information is key (MRI, ultrasound) [69, 70].

Data and Generalization: The availability and variability of training data differ by modality [71]. This directly impacts how well a given architecture can realize its potential performance [72, 73].

Ultimately, the physical limitations of an image can be amplified by the constraints of an AI model, or in some cases, complemented by them, as seen with ultrasound. This integrated perspective suggests that the problem should be reframed from "which modality is best?" to "which model is best suited for the unique characteristics of each modality?"

5 Discussion

5.1 Limitations of Current Research

A review of existing literature reveals several limitations:

Methodological Bias: The vast majority (approx. 98%) of medical imaging AI studies are retrospective [74], with a scarcity of prospective studies or randomized controlled trials [75]. This introduces potential bias and may not reflect real-world clinical effectiveness.

Reporting and Publication Bias: Many studies claim AI performance is equivalent or superior to clinicians [76, 77], yet less than half (38%) conduct direct comparative evaluations. This suggests a tendency to publish positive results and potentially overstate claims.

Lack of Standardization and Reproducibility: Many AI studies fail to adhere to reporting guidelines like TRIPOD [78], omitting crucial details about data pre-processing and model specifics. This raises concerns about the reproducibility and reliability of the findings.

5.2 Clinical Significance and Practical Implications

Despite these limitations, the tangible benefits of AI in the clinical setting are undeniable:

Improved Workflow Efficiency: AI integration has been shown to reduce image interpretation time by an average of 27.20% and workload by 58.48% [79], alleviating the burden on radiologists.

Enhanced Diagnostic Accuracy and Consistency: AI assistance can significantly improve the performance of less experienced physicians, with one study showing a 24% increase in sensitivity [80], helping to standardize the quality of care.

Patient Safety and Cost Reduction: AI-driven techniques enable significant reductions in radiation dose (by over 50%) and contrast agent use (by 80-90%), enhancing patient safety while also reducing healthcare costs [81, 82].

5.3 A Practical Solution: The Hybrid Workflow (Hypothesis 3)

Based on our analysis, we propose a hybrid diagnostic workflow that strategically combines AI systems with complementary strengths. This multi-stage decision-making process is designed to maximize the advantages of each imaging modality.

Stage 1 – Broad Screening: In the initial phase, low-cost, high-sensitivity AI modalities like X-ray or ultrasound are used. The focus is on capturing any potential abnormalities and filtering out the majority of normal cases.

Stage 2 – Precision Diagnosis: Cases flagged in Stage 1 proceed to high-resolution, high-specificity modalities like CT or MRI. Here, a second AI system focuses on reducing false positives and accurately characterizing lesions for definitive diagnosis and treatment planning.

Stage 3 – Integrated Decision: A clinician makes the final judgment by integrating the results from both stages. This multi-modal ensemble approach has been reported to improve accuracy by over 17% compared to single-modality models [83, 84].

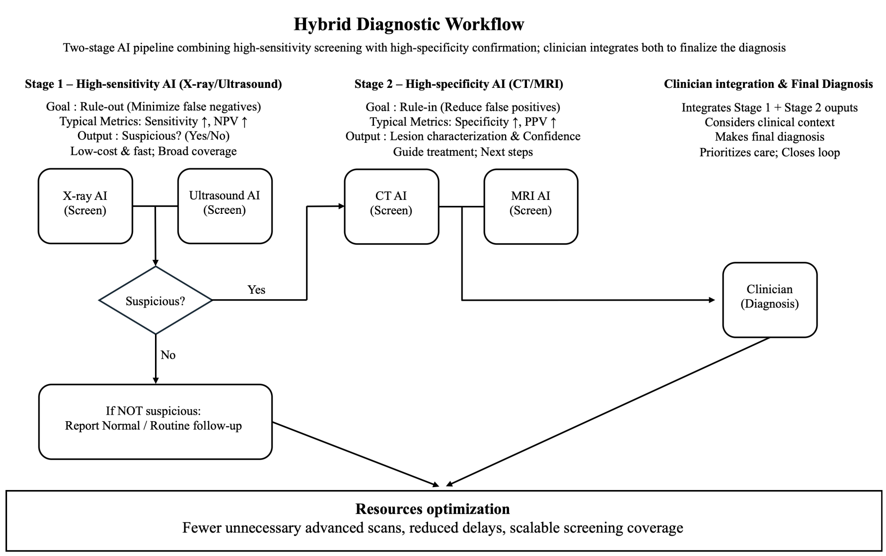

Figure 1: a proposed hybrid diagnostic workflow.
Stage 1 uses high-sensitivity AI (X-ray/ultrasound) for broad screening, and cases flagged as suspicious proceed to Stage 2 for precision diagnosis with high-specificity AI (CT/MRI). A clinician integrates both stages to make the final diagnosis, optimizing resource utilization and reducing diagnostic delays.

5.4 Future Research Directions

Future research should focus on enhancing the reliability, efficiency, and applicability of medical imaging AI:

Explainable AI (XAI): Developing more intuitive and robust XAI techniques (e.g., SHAP, LIME, Grad-CAM) is crucial to overcome the "black box" nature of deep learning and build clinical trust.

Foundation Models and Multi-modal AI: The development of large-scale foundation models pre-trained on millions of medical images could mitigate data scarcity issues. Furthermore, multi-modal AI that integrates imaging with clinical text (e.g., radiology reports) holds great promise for comprehensive clinical decision support.

Real-time Adaptive Systems: AI systems that can adapt in real-time to patient-specific characteristics or intra-procedural events are needed. This requires advancements in edge AI and on-device learning.

Sustainable and Accessible Technology: Pairing AI with sustainable hardware, such as helium-free MRI and portable ultrasound/X-ray devices, can help bridge global healthcare disparities.

Data Sharing and Governance: Privacy-preserving techniques like Federated Learning are essential for collaborative research. Establishing standardized data formats and performance benchmarks is also a key task for the research community and regulatory bodies.

6 Conclusion

This study has systematically analyzed the performance discrepancies of AI across different medical imaging modalities, diagnosing their causes and proposing strategic solutions.

a. Empirical Confirmation of Performance Gaps: We confirmed that AI performance varies significantly by modality, with ultrasound-based AI showing the highest performance (AUROC 0.94), followed by CT/MRI ($\tilde{0}$.82), while X-ray exhibits greater variability.

b. A Complex Interplay of Causes: The performance gap results from a complex interaction between the physical properties of the images and the structural limitations of current AI architectures, particularly the constraints of CNNs in handling global and spatio-temporal information.

c. The Promise of a Hybrid Workflow: A hybrid approach that strategically combines the different strengths of modality-specific AIs (high-sensitivity for screening, high-specificity for confirmation) was proposed as a practical and effective solution.

d. Demonstrated Clinical Value: AI integration has proven its value by improving workflow efficiency (27% faster interpretation), enhancing diagnostic accuracy (12% sensitivity increase), and improving patient safety (over 50% radiation dose reduction).

The core contribution of this work is the systematic framing of the AI performance gap through a logical progression from **phenomenon** → **cause** → **solution**, culminating in the proposal of a practical hybrid workflow. The future of medical AI lies not in perfecting a single model for one modality, but in developing an integrated and collaborative ecosystem where AI, clinicians, and diverse data sources work in concert. Achieving this vision will require continued technological innovation alongside concerted efforts in clinician education, regulatory adaptation, and ethical governance.

Agents4Science AI Involvement Checklist

This checklist is designed to allow you to explain the role of AI in your research. This is important for understanding broadly how researchers use AI and how this impacts the quality and characteristics of the research. **Do not remove the checklist! Papers not including the checklist will be desk rejected.** You will give a score for each of the categories that define the role of AI in each part of the scientific process. The scores are as follows:

- **[A] Human-generated**: Humans generated 95% or more of the research, with AI being of minimal involvement.
- **[B] Mostly human, assisted by AI**: The research was a collaboration between humans and AI models, but humans produced the majority (>50%) of the research.
- **[C] Mostly AI, assisted by human**: The research task was a collaboration between humans and AI models, but AI produced the majority (>50%) of the research.
- **[D] AI-generated**: AI performed over 95% of the research. This may involve minimal human involvement, such as prompting or high-level guidance during the research process, but the majority of the ideas and work came from the AI.

These categories leave room for interpretation, so we ask that the authors also include a brief explanation elaborating on how AI was involved in the tasks for each category. Please keep your explanation to less than 150 words.

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[B]**

   Explanation:The initial ideas for the hypotheses were proposed by human researchers, while the AI evaluated their validity through in-depth research, providing assessments of feasibility along with supporting scholarly papers.

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: **[B]**

   Explanation: Human authors lack any knowledge in computer science or engineering, rendering them unable to comprehend the experimental designs proposed by the AI. Consequently, the human authors suggested the experimental designs and research methods, which the AI subsequently verified.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[A]**

   Explanation: As the AI did not directly perform coding or data analysis in this paper, interpretations generated by the AI are not included.

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[C]**

   Explanation: Since some human authors are not native English speakers, AI translation features were extensively utilized. The human authors continually imposed various requirements on the text generated by the AI. For instance, "In our view, our expressions more accurately reflect our intentions than yours. Therefore, we have revised your expressions and sentences."

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

Description: In conducting this research in collaboration with AI, we conclude that the ability to create something from nothing remains a distant goal. Nevertheless, when humans devoid of specialized expertise propose an idea, the AI employs all available means to evaluate it by presenting appropriate rationales. We are confident that this represents a significant advancement in the scientific community, enabling unprecedented innovations through a single idea, without the need for advanced intelligence or knowledge.

Agents4Science Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **Papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers and area chairs. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given. In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's main contributions: the analysis of AI performance discrepancy across modalities, the investigation of its causes, and the proposal of a hybrid workflow as a solution. These claims are consistently supported by the literature review and discussion in the main body.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The "5.1 Limitations of Current Research" subsection within the Discussion section explicitly addresses the limitations of the existing literature on which this review is based, such as the predominance of retrospective studies and potential publication bias.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.

- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper is a review and perspective article; it does not present new theoretical results or mathematical proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: his paper does not contain original experimental results. All claims are based on publicly available, cited literature.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: This paper does not involve original code or data, as it is a literature review.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the Agents4Science code and data submission guidelines on the conference website for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: This paper reviews existing studies and does not present its own experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: When citing results from other studies that support our main claims, we have included statistical information such as 95% confidence intervals (CI) and AUROC values where available in the source literature.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: This paper does not report on new experiments, so compute resources are not applicable.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

Answer: [Yes]

Justification: This research is based on a review of existing, publicly available literature and adheres to standard principles of academic integrity and responsible conduct.

Guidelines:

- The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses positive societal impacts in the "5.2 Clinical Significance" section (e.g., improved efficiency, patient safety) and touches upon negative aspects and ethical considerations in the "Regulatory and Ethical Considerations" subsection (e.g., issues of bias, privacy, and legal responsibility).

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies.

References

[1] Samuel L Brady, Andrew T Trout, Elanchezhian Somasundaram, Christopher G Anton, Yinan Li, and Jonathan R Dillman. Improving Image Quality and Reducing Radiation Dose for Pediatric CT by Using Deep Learning Reconstruction. *Radiology*, 298(1):180–188, 2021. doi: 10.1148/radiol.2020202317. URL https://doi.org/10.1148/radiol.2020202317.

[2] Lennart R Koetzier, Domenico Mastrodicasa, Timothy P Szczykutowicz, Niels R van der Werf, Adam S Wang, Veit Sandfort, Aart J van der Molen, Dominik Fleischmann, and Martin J Willemink. Deep Learning Image Reconstruction for CT: Technical Principles and Clinical Prospects. *Radiology*, 306(3):e221257, 2023. doi: 10.1148/radiol.221257. URL https://doi.org/10.1148/radiol.221257.

[3] Gaspard Ludes, Mickael Ohana, Aissam Labani, Nicolas Meyer, Sébastien Moliére, and Catherine Roy. Impact of a reduced iodine load with deep learning reconstruction on abdominal MDCT. *Medicine*, 102(35), 2023. URL https://journals.lww.com/md-journal/fulltext/2023/09010/impact_of_a_reduced_iodine_load_with_deep_learning.72.aspx.

[4] Achille Mileto, Lifeng Yu, Jonathan W Revels, Serageldin Kamel, Mostafa A Shehata, Juan J Ibarra-Rovira, Vincenzo K Wong, Alicia M Roman-Colon, Jeong Min Lee, Khaled M Elsayes, and Corey T Jensen. State-of-the-Art Deep Learning CT Reconstruction Algorithms in Abdominal Imaging. *RadioGraphics*, 44(12):e240095, 11 2024. ISSN 0271-5333. doi: 10.1148/rg.240095. URL `https://doi.org/10.1148/rg.240095`.

[5] I Ahmed, A Chehri, G Jeon, and F Piccialli. Automated Pulmonary Nodule Classification and Detection Using Deep Learning Architectures. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(4):2445–2456, 2023. ISSN 1557-9964. doi: 10.1109/TCBB.2022.3192139.

[6] Pierre P Massion, Sanja Antic, Sarim Ather, Carlos Arteta, Jan Brabec, Heidi Chen, Jerome Declerck, David Dufek, William Hickes, Timor Kadir, Jonas Kunst, Bennett A Landman, Reginald F Munden, Petr Novotny, Heiko Peschl, Lyndsey C Pickup, Catarina Santos, Gary T Smith, Ambika Talwar, and Fergus Gleeson. Assessing the Accuracy of a Deep Learning Method to Risk Stratify Indeterminate Pulmonary Nodules. *American Journal of Respiratory and Critical Care Medicine*, 202(2):241–249, 4 2020. ISSN 1073-449X. doi: 10.1164/rccm.201903-0505OC. URL `https://doi.org/10.1164/rccm.201903-0505OC`.

[7] Qi Dou, Hao Chen, Lequan Yu, Lei Zhao, Jing Qin, Defeng Wang, Vincent C.T. Mok, Lin Shi, and Pheng Ann Heng. Automatic Detection of Cerebral Microbleeds from MR Images via 3D Convolutional Neural Networks. *IEEE Transactions on Medical Imaging*, 35(5):1182–1195, 5 2016. ISSN 1558254X. doi: 10.1109/TMI.2016.2528129.

[8] Yin Dai, Yifan Gao, and Fayu Liu. TransMed: Transformers Advance Multi-Modal Medical Image Classification. *Diagnostics*, 11(8), 2021. ISSN 2075-4418. doi: 10.3390/diagnostics11081384. URL `https://www.mdpi.com/2075-4418/11/8/1384`.

[9] Chunhui Zhao, Boao Qin, Shou Feng, Wenxiang Zhu, Weiwei Sun, Wei Li, and Xiuping Jia. Hyperspectral Image Classification With Multi-Attention Transformer and Adaptive Superpixel Segmentation-Based Active Learning. *IEEE Transactions on Image Processing*, 32:3606–3621, 2023. doi: 10.1109/TIP.2023.3287738.

[10] Mohammed Aloraini, Asma Khan, Suliman Aladhadh, Shabana Habib, Mohammed F Alsharekh, and Muhammad Islam. Combining the Transformer and Convolution for Effective Brain Tumor Classification Using MRI Images. *Applied Sciences*, 2023. URL `https://api.semanticscholar.org/CorpusID:257556304`.

[11] Md. Mahfuz Ahmed, Md. Maruf Hossain, Md. Rakibul Islam, Md. Shahin Ali, Abdullah Al Noman Nafi, Md. Faisal Ahmed, Kazi Mowdud Ahmed, Md. Sipon Miah, Md. Mahbubur Rahman, Mingbo Niu, and Md. Khairul Islam. Brain tumor detection and classification in MRI using hybrid ViT and GRU model with explainable AI in Southern Bangladesh. *Scientific Reports*, 14(1):22797, 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-71893-3. URL `https://doi.org/10.1038/s41598-024-71893-3`.

[12] Marta Mirkov and Member Ieee Ana Gavrovska. Application of Bayes and knn classifiers in tumor detection from brain MRI images. 2022. URL `https://api.semanticscholar.org/CorpusID:254588068`.

[13] Eric A Walker, Michael E Fenton, Joel S Salesky, and Mark D Murphey. Magnetic Resonance Imaging of Benign Soft Tissue Neoplasms in Adults. *Radiologic Clinics*, 49(6):1197–1217, 11 2011. ISSN 0033-8389. doi: 10.1016/j.rcl.2011.07.007. URL `https://doi.org/10.1016/j.rcl.2011.07.007`.

[14] A M Aisen, W Martel, E M Braunstein, K I McMillin, W A Phillips, and T F Kling. MRI and CT evaluation of primary bone and soft-tissue tumors. *American Journal of Roentgenology*, 146(4):749–756, 1986. doi: 10.2214/ajr.146.4.749. URL `https://doi.org/10.2214/ajr.146.4.749`.

[15] Heang-Ping Chan, Ravi K Samala, Lubomir M Hadjiiski, and Chuan Zhou. Deep Learning in Medical Image Analysis. In Gobert Lee and Hiroshi Fujita, editors, *Deep Learning in Medical Image Analysis : Challenges and Applications*, pages 3–21. Springer International Publishing, Cham, 2020. ISBN 978-3-030-33128-3. doi: 10.1007/978-3-030-33128-3{_}1. URL `https://doi.org/10.1007/978-3-030-33128-3_1`.

[16] Panagiotis Papadimitroulas, Lennart Brocki, Neo Christopher Chung, Wistan Marchadour, Franck Vermet, Laurent Gaubert, Vasilis Eleftheriadis, Dimitris Plachouris, Dimitris Visvikis, George C Kagadis, and Mathieu Hatt. Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization. *Physica Medica: European Journal of Medical Physics*, 83:108–121, 3 2021. ISSN 1120-1797. doi: 10.1016/j.ejmp.2021.03.009. URL `https://doi.org/10.1016/j.ejmp.2021.03.009`.

[17] C Kishor Kumar Reddy, Pulakurthi Anaghaa Reddy, Himaja Janapati, Basem Assiri, Mohammed Shuaib, Shadab Alam, and Abdullah Sheneamer. A fine-tuned vision transformer based enhanced multi-class brain tumor classification using MRI scan imagery. *Frontiers in Oncology*, Volume 14 - 2024, 2024. ISSN 2234-943X. doi: 10.3389/fonc.2024.1400341. URL `https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2024.1400341`.

[18] Sara Tehsin, Inzamam Mashood Nasir, and Robertas Damaševičius. GATransformer: A Graph Attention Network-Based Transformer Model to Generate Explainable Attentions for Brain Tumor Detection. *Algorithms*, 18(2), 2025. ISSN 1999-4893. doi: 10.3390/a18020089. URL `https://www.mdpi.com/1999-4893/18/2/89`.

[19] Martin Bendszus, Andrea Laghi, Josep Munuera, Lawrence N Tanenbaum, Bachir Taouli, and Harriet C Thoeny. MRI Gadolinium-Based Contrast Media: Meeting Radiological, Clinical, and Environmental Needs. *Journal of Magnetic Resonance Imaging*, 60(5):1774–1785, 2024. doi: https://doi.org/10.1002/jmri.29181. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.29181`.

[20] Srivathsa Pasumarthi, Jonathan I Tamir, Søren Christensen, Greg Zaharchuk, Tao Zhang, and Enhao Gong. A generic deep learning model for reduced gadolinium dose in contrast-enhanced brain MRI. *Magnetic Resonance in Medicine*, 86:1687 – 1700, 2021. URL `https://api.semanticscholar.org/CorpusID:233461257`.

[21] Dana J Lin, Patricia M Johnson, Florian Knoll, and Yvonne W Lui. Artificial Intelligence for MR Image Reconstruction: An Overview for Clinicians. *Journal of Magnetic Resonance Imaging*, 53(4):1015–1028, 2021. doi: https://doi.org/10.1002/jmri.27078. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.27078`.

[22] Yutong Chen, Carola-Bibiane Schönlieb, Pietro Liò, Tim Leiner, Pier Luigi Dragotti, Ge Wang, Daniel Rueckert, David Firmin, and Guang Yang. AI-Based Reconstruction for Fast MRI—A Systematic Review and Meta-Analysis. *Proceedings of the IEEE*, 110(2):224–245, 2022. doi: 10.1109/JPROC.2022.3141367.

[23] Hayden P Baker, Sarthak Aggarwal, Senthooran Kalidoss, Matthew Hess, Rex Haydon, and Jason A Strelzow. Diagnostic accuracy of ChatGPT-4 in orthopedic oncology: a comparative study with residents. *The Knee*, 55:153–160, 2025. ISSN 0968-0160. doi: https://doi.org/10.1016/j.knee.2025.04.004. URL `https://www.sciencedirect.com/science/article/pii/S0968016025000766`.

[24] Alperen Elek, Duygu Doğa Ekizalioğlu, and Ezgi Güler. Evaluating Microsoft Bing with ChatGPT-4 for the assessment of abdominal computed tomography and magnetic resonance images. *Diagnostic and Interventional Radiology*, 31:196 – 205, 2024. URL `https://api.semanticscholar.org/CorpusID:271896071`.

[25] Junting Zhao, Zhaohu Xing, Zhihao Chen, Liang Wan, Tong Han, H Fu, and Lei Zhu. Uncertainty-Aware Multi-Dimensional Mutual Learning for Brain and Brain Tumor Segmentation. *IEEE Journal of Biomedical and Health Informatics*, 27:4362–4372, 2023. URL `https://api.semanticscholar.org/CorpusID:258565652`.

[26] Zhiliang Peng, Zonghao Guo, Wei Huang, Yaowei Wang, Lingxi Xie, Jianbin Jiao, Qi Tian, and Qixiang Ye. Conformer: Local Features Coupling Global Representations for Recognition and Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:9454–9468, 2023. URL `https://api.semanticscholar.org/CorpusID:256683188`.

[27] Ali Hatamizadeh Nvidia, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, Nvidia Daguang, and Xu Nvidia. UNETR: Transformers for 3D Medical Image Segmentation. Technical report. URL https://monai.io/research/unetr.

[28] Konstantinos Kamnitsas, Christian Ledig, Virginia F J Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2016.10.004. URL https://www.sciencedirect.com/science/article/pii/S1361841516301839.

[29] Zanyar HajiEsmailPoor, Peyman Tabnak, Behzad Baradaran, Fariba Pashazadeh, and Leili Aghebati-Maleki. Diagnostic performance of CT scan–based radiomics for prediction of lymph node metastasis in gastric cancer: a systematic review and meta-analysis. *Frontiers in Oncology*, Volume 13 - 2023, 2023. ISSN 2234-943X. doi: 10.3389/fonc.2023.1185663. URL https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2023.1185663.

[30] Cenji Yu, Chidinma P Anakwenze, Yao Zhao, Rachael M Martin, Ethan B Ludmir, Joshua S.Niedzielski, Asad Qureshi, Prajnan Das, Emma B Holliday, Ann C Raldow, Callistus M Nguyen, Raymond P Mumme, Tucker J Netherton, Dong Joo Rhee, Skylar S Gay, Jinzhong Yang, Laurence E Court, and Carlos E Cardenas. Multi-organ segmentation of abdominal structures from non-contrast and contrast enhanced CT images. *Scientific Reports*, 12(1):19093, 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-21206-3. URL https://doi.org/10.1038/s41598-022-21206-3.

[31] Aidan Boyd, Zezhong Ye, Sanjay P Prabhu, Michael C Tjong, Yining Zha, Anna Zapaishchykova, Sridhar Vajapeyam, Paul J Catalano, Hasaan Hayat, Rishi Chopra, Kevin X Liu, Ali Nabavizadeh, Adam C Resnick, Sabine Mueller, Daphne A Haas-Kogan, Hugo J W L Aerts, Tina Y Poussaint, and Benjamin H Kann. Stepwise Transfer Learning for Expert-level Pediatric Brain Tumor MRI Segmentation in a Limited Data Scenario. *Radiology: Artificial Intelligence*, 6(4):e230254, 2024. doi: 10.1148/ryai.230254. URL https://doi.org/10.1148/ryai.230254.

[32] Arnaldo Stanzione, Renato Cuocolo, Lorenzo Ugga, Francesco Verde, Valeria Romeo, Arturo Brunetti, and Simone Maurea. Oncologic Imaging and Radiomics: A Walkthrough Review of Methodological Challenges. *Cancers*, 14(19), 2022. ISSN 2072-6694. doi: 10.3390/cancers14194871. URL https://www.mdpi.com/2072-6694/14/19/4871.

[33] Jong Seok Ahn, Shadi Ebrahimian, Shaunagh McDermott, Sanghyup Lee, Laura Naccarato, John F Di Capua, Markus Y Wu, Eric W Zhang, Victorine Muse, Benjamin Miller, Farid Sabzalipour, Bernardo C Bizzo, Keith J Dreyer, Parisa Kaviani, Subba R Digumarthy, and Mannudeep K Kalra. Association of Artificial Intelligence–Aided Chest Radiograph Interpretation With Reader Performance and Efficiency. *JAMA Network Open*, 5(8):e2229289–e2229289, 8 2022. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2022.29289. URL https://doi.org/10.1001/jamanetworkopen.2022.29289.

[34] Michaël Chassé and Dean A Fergusson. Diagnostic Accuracy Studies. *Seminars in Nuclear Medicine*, 49(2):87–93, 2019. ISSN 0001-2998. doi: https://doi.org/10.1053/j.semnuclmed.2018.11.005. URL https://www.sciencedirect.com/science/article/pii/S0001299818300941.

[35] David Hua, Khang Nguyen, Neysa Petrina, Noel Young, Jin-Gun Cho, Adeline Yap, and Simon K Poon. Benchmarking the diagnostic test accuracy of certified AI products for screening pulmonary tuberculosis in digital chest radiographs: Preliminary evidence from a rapid review and meta-analysis. *International Journal of Medical Informatics*, 177:105159, 2023. ISSN 1386-5056. doi: https://doi.org/10.1016/j.ijmedinf.2023.105159. URL https://www.sciencedirect.com/science/article/pii/S1386505623001776.

[36] Peng Tang, Qiaokang Liang, Xintong Yan, Shao Xiang, and Dan Zhang. GP-CNN-DTEL: Global-Part CNN Model With Data-Transformed Ensemble Learning for Skin Lesion Classification. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2870–2882, 2020. doi: 10.1109/JBHI.2020.2977013.

[37] Julius Husarek, Silvan Hess, Sam Razaeian, Thomas D Ruder, Stephan Sehmisch, Martin Müller, and Emmanouil Liodakis. Artificial intelligence in commercial fracture detection products: a systematic review and meta-analysis of diagnostic test accuracy. *Scientific Reports*, 14(1):23053, 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-73058-8. URL https://doi.org/10.1038/s41598-024-73058-8.

[38] Cathy Ong Ly, Balagopal Unnikrishnan, Tony Tadic, Tirth Patel, Joe Duhamel, Sonja Kandel, Yasbanoo Moayedi, Michael Brudno, Andrew Hope, Heather Ross, and Chris McIntosh. Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data. *npj Digital Medicine*, 7(1):124, 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01118-4. URL https://doi.org/10.1038/s41746-024-01118-4.

[39] Yuanbin Chen, Tao Wang, Hui Tang, Longxuan Zhao, Xinlin Zhang, Tao Tan, Qinquan Gao, Min Du, and Tong Tong. CoTrFuse: a novel framework by fusing CNN and transformer for medical image segmentation. *Physics in Medicine & Biology*, 68(17):175027, 8 2023. doi: 10.1088/1361-6560/acede8. URL https://dx.doi.org/10.1088/1361-6560/acede8.

[40] S Suganyadevi, V Seethalakshmi, and K Balasamy. A review on deep learning in medical image analysis. *International Journal of Multimedia Information Retrieval*, 11(1):19–38, 2022. ISSN 2192-662X. doi: 10.1007/s13735-021-00218-1. URL https://doi.org/10.1007/s13735-021-00218-1.

[41] Lin Ma, Liqiong Huang, Yan Chen, Lei Zhang, Dunli Nie, Wenjing He, and Xiaoxue Qi. AI diagnostic performance based on multiple imaging modalities for ovarian tumor: A systematic review and meta-analysis. *Frontiers in Oncology*, 13, 2023. URL https://api.semanticscholar.org/CorpusID:258241542.

[42] Asaf Raza, Naeem Ullah, Javed Ali Khan, Muhammad Assam, Antonella Guzzo, and Hanan Aljuaid. DeepBreastCancerNet: A Novel Deep Learning Model for Breast Cancer Detection Using Ultrasound Images. *Applied Sciences*, 13(4), 2023. ISSN 2076-3417. doi: 10.3390/app13042082. URL https://www.mdpi.com/2076-3417/13/4/2082.

[43] Ji Lin, Chuang Gan, Kuan Wang, and Song Han. TSM: Temporal Shift Module for Efficient and Scalable Video Understanding on Edge Devices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2760–2774, 2022. doi: 10.1109/TPAMI.2020.3029799.

[44] Ji Lin MIT, Chuang Gan, and Song Han MIT. TSM: Temporal Shift Module for Efficient Video Understanding. Technical report. URL https://github.

[45] Hiren Mewada, Jawad F Al-Asad, Himanshu Patel, and Nayeemuddin Mohammed. Leveraging Spatial and Temporal Features using CNN-LSTM for Improved Bone Fracture Classification from X-ray Images. In *2024 6th International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*, pages 1–5, 2024. doi: 10.1109/ISAECT64333.2024.10799900.

[46] Pallabi Saikia, Dhwani Dholaria, Priyanka Yadav, Vaidehi Patel, and Mohendra Roy. A Hybrid CNN-LSTM model for Video Deepfake Detection by Leveraging Optical Flow Features. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2022. doi: 10.1109/IJCNN55064.2022.9892905.

[47] Zhizhong Huang, Junping Zhang, Yi Zhang, and Hongming Shan. DU-GAN: Generative Adversarial Networks With Dual-Domain U-Net-Based Discriminators for Low-Dose CT Denoising. *IEEE Transactions on Instrumentation and Measurement*, 71:1–12, 2022. doi: 10.1109/TIM.2021.3128703.

[48] Mufeng Geng, Xiangxi Meng, Jiangyuan Yu, Lei Zhu, Lujia Jin, Zhe Jiang, Bin Qiu, Hui Li, Hanjing Kong, Jianmin Yuan, Kun Yang, Hongming Shan, Hongbin Han, Zhi Yang, Qiushi Ren, and Yanye Lu. Content-Noise Complementary Learning for Medical Image Denoising. *IEEE Transactions on Medical Imaging*, 41:407–419, 2021. URL https://api.semanticscholar.org/CorpusID:237547546.

[49] Ahmad Chaddad, Qizong Lu, Jiali Li, Yousef Katib, Reem Kateb, Camel Tanougast, Ahmed Bouridane, and Ahmed Abdulkadir. Explainable, Domain-Adaptive, and Federated Artificial Intelligence in Medicine. *IEEE/CAA Journal of Automatica Sinica*, 10(4):859–876, 2023. doi: 10.1109/JAS.2023.123123.

[50] Seong Ho Park and Kyunghwa Han. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. *Radiology*, 286(3):800–809, 2018. doi: 10.1148/radiol.2017171920. URL `https://doi.org/10.1148/radiol.2017171920`.

[51] F Maes, A Collignon, D Vandermeulen, G Marchal, and P Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16(2):187–198, 1997. doi: 10.1109/42.563664.

[52] B W Raaymakers, J J W Lagendijk, J Overweg, J G M Kok, A J E Raaijmakers, E M Kerkhof, R W van der Put, I Meijsing, S P M Crijns, F Benedosso, M van Vulpen, C H W de Graaff, J Allen, and K J Brown. Integrating a 1.5 T MRI scanner with a 6 MV accelerator: proof of concept. *Physics in Medicine & Biology*, 54(12):N229, 5 2009. doi: 10.1088/0031-9155/54/12/N01. URL `https://dx.doi.org/10.1088/0031-9155/54/12/N01`.

[53] Bob D de Vos, Jelmer M Wolterink, Pim A de Jong, Tim Leiner, Max A Viergever, and Ivana Išgum. ConvNet-Based Localization of Anatomical Structures in 3-D Medical Images. *IEEE Transactions on Medical Imaging*, 36(7):1470–1481, 2017. doi: 10.1109/TMI.2017.2673121.

[54] Wonseok Choi, Byullee Park, Seongwook Choi, Donghyeon Oh, Jongbeom Kim, and Chulhong Kim. Recent Advances in Contrast-Enhanced Photoacoustic Imaging: Overcoming the Physical and Practical Challenges. *Chemical Reviews*, 123(11):7379–7419, 6 2023. ISSN 0009-2665. doi: 10.1021/acs.chemrev.2c00627. URL `https://doi.org/10.1021/acs.chemrev.2c00627`.

[55] Tilman Donath, Franz Pfeiffer, Oliver Bunk, Christian Grünzweig, Eckhard Hempel, Stefan Popescu, Peter Vock, and Christian David. Toward Clinical X-ray Phase-Contrast CT: Demonstration of Enhanced Soft-Tissue Contrast in Human Specimen. *Investigative Radiology*, 45(7), 2010. URL `https://journals.lww.com/investigativeradiology/fulltext/2010/07000/toward_clinical_x_ray_phase_contrast_ct_.12.aspx`.

[56] Kate Sanders, Reno Kriz, Anqi Liu, and Benjamin Van Durme. Ambiguous Images With Human Judgments for Robust Visual Event Classification. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 2637–2650. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/11e3e0f1b29dcd31bd0952bfc1357f68-Paper-Datasets_and_Benchmarks.pdf`.

[57] Yury Rusinovich, Volha Rusinovich, and Markus Doss. Bionic Hand Control with Real-Time B-Mode Ultrasound Web AI Vision. *Web3 Journal: ML in Health Science*, 2025. URL `https://api.semanticscholar.org/CorpusID:277599004`.

[58] D R Sarvamangala and Raghavendra V Kulkarni. Convolutional neural networks in medical image understanding: a survey. *Evolutionary Intelligence*, 15(1):1–22, 2022. ISSN 1864-5917. doi: 10.1007/s12065-020-00540-3. URL `https://doi.org/10.1007/s12065-020-00540-3`.

[59] Ahmad Waleed Salehi, Shakir Khan, Gaurav Gupta, Bayan Ibrahimm Alabduallah, Abrar Almjally, Hadeel Alsolai, Tamanna Siddiqui, and Adel Mellit. A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope. *Sustainability*, 15 (7), 2023. ISSN 2071-1050. doi: 10.3390/su15075930. URL `https://www.mdpi.com/2071-1050/15/7/5930`.

[60] R Archana and P S Eliahim Jeevaraj. Deep learning models for digital image processing: a review. *Artificial Intelligence Review*, 57(1):11, 2024. ISSN 1573-7462. doi: 10.1007/s10462-023-10631-z. URL `https://doi.org/10.1007/s10462-023-10631-z`.

[61] Zihan Li, Yuan Zheng, Dandan Shan, Shuzhou Yang, Qingde Li, Beizhan Wang, Yuanting Zhang, Qingqi Hong, and Dinggang Shen. ScribFormer: Transformer Makes CNN Work Better for Scribble-Based Medical Image Segmentation. *IEEE Transactions on Medical Imaging*, 43 (6):2254–2265, 2024. doi: 10.1109/TMI.2024.3363190.

[62] Md Shofiqul Islam, Mst Sunjida Sultana, Uttam Kumar Roy, and Jubayer Al Mahmud. A review on Video Classification with Methods, Findings, Performance, Challenges, Limitations and Future Work. *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, 6(2):47, 1 2021. ISSN 2338-3070. doi: 10.26555/jiteki.v6i2.18978.

[63] Shutian Zhao, Fan Xiao, James F Griffith, Ruokun Li, and Weitian Chen. Denoising of volumetric magnetic resonance imaging using multi-channel three-dimensional convolutional neural network with applications on fast spin echo acquisitions. *Quantitative Imaging in Medicine and Surgery*, 14:6517 – 6530, 2023. URL https://api.semanticscholar.org/CorpusID:272188162.

[64] Guangle Yao, Tao Lei, and Jiandan Zhong. A review of Convolutional-Neural-Network-based action recognition. *Pattern Recognition Letters*, 118:14–22, 2019. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2018.05.018. URL https://www.sciencedirect.com/science/article/pii/S0167865518302058.

[65] Satya P Singh, Lipo Wang, Sukrit Gupta, Haveesh Goli, Parasuraman Padmanabhan, and Balázs Gulyás. 3D Deep Learning on Medical Images: A Review. *Sensors*, 20(18), 2020. ISSN 1424-8220. doi: 10.3390/s20185097. URL https://www.mdpi.com/1424-8220/20/18/5097.

[66] Pengfei Gu, Yejia Zhang, Chaoli Wang, and Da Chen. ConvFormer: Combining CNN and Transformer for Medical Image Segmentation. *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2022. URL https://api.semanticscholar.org/CorpusID:253553289.

[67] Lingting Zhu, Yizheng Chen, Lianli Liu, Lei Xing, and Lequan Yu. Multi-Sensor Learning Enables Information Transfer Across Different Sensory Data and Augments Multi-Modality Imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):288–304, 2025. doi: 10.1109/TPAMI.2024.3465649.

[68] Li Yan, Qing Li, Kang Fu, Xiaodong Zhou, and Kai Zhang. Progress in the Application of Artificial Intelligence in Ultrasound-Assisted Medical Diagnosis. *Bioengineering*, 12(3), 2025. ISSN 2306-5354. doi: 10.3390/bioengineering12030288. URL https://www.mdpi.com/2306-5354/12/3/288.

[69] C Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and Multi-view CNNs for Object Classification on 3D Data. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5648–5656, 2016. URL https://api.semanticscholar.org/CorpusID:1009127.

[70] Juezhao Yu, Bohan Yang, Jing Wang, Joseph K Leader, David O Wilson, and Jiantao Pu. 2D CNN versus 3D CNN for false-positive reduction in lung cancer screening. *Journal of Medical Imaging*, 7(5):51202, 10 2020. doi: 10.1117/1.jmi.7.5.051202. URL https://lens.org/041-586-834-140-199.

[71] Félix Renard, Soulaimane Guedria, Noel De Palma, and Nicolas Vuillerme. Variability and reproducibility in deep learning for medical image segmentation. *Scientific Reports*, 10(1): 13724, 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-69920-0. URL https://doi.org/10.1038/s41598-020-69920-0.

[72] Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *npj Digital Medicine*, 5(1):149, 2022. ISSN 2398-6352. doi: 10.1038/s41746-022-00689-4. URL https://doi.org/10.1038/s41746-022-00689-4.

[73] Natália Alves, Joeran S Bosma, Kiran V Venkadesh, Colin Jacobs, Zaigham Saghir, Maarten de Rooij, John Hermans, and Henkjan Huisman. Prediction Variability to Identify Reduced AI Performance in Cancer Diagnosis at MRI and CT. *Radiology*, 308(3):e230275, 2023. doi: 10.1148/radiol.230275. URL https://doi.org/10.1148/radiol.230275.

[74] Myura Nagendran, Yang Chen, Christopher A Lovejoy, Anthony C Gordon, Matthieu Komorowski, Hugh Harvey, Eric J Topol, John P A Ioannidis, Gary S Collins, and Mahiben Maruthappu. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*, 368, 2020. doi: 10.1136/bmj.m689. URL https://www.bmj.com/content/368/bmj.m689.

[75] Ryan Han, Julián N Acosta, Zahra Shakeri, John P A Ioannidis, Eric J Topol, and Pranav Rajpurkar. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *The Lancet Digital Health*, 6(5):e367–e373, 5 2024. ISSN 2589-7500. doi: 10.1016/S2589-7500(24)00047-5. URL `https://doi.org/10.1016/S2589-7500(24)00047-5`.

[76] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019. ISSN 1546-170X. doi: 10.1038/s41591-018-0300-7. URL `https://doi.org/10.1038/s41591-018-0300-7`.

[77] Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R Ledsam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, and Alastair K Denniston. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, 1(6):e271–e297, 10 2019. ISSN 2589-7500. doi: 10.1016/S2589-7500(19)30123-2. URL `https://doi.org/10.1016/S2589-7500(19)30123-2`.

[78] Gary S Collins, Paula Dhiman, Constanza L Andaur Navarro, Jie Ma, Lotty Hooft, Johannes B Reitsma, Patricia Logullo, Andrew L Beam, Lily Peng, Ben Van Calster, Maarten van Smeden, Richard D Riley, and Karel G M Moons. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*, 11(7), 2021. ISSN 2044-6055. doi: 10.1136/bmjopen-2020-048008. URL `https://bmjopen.bmj.com/content/11/7/e048008`.

[79] Mingyang Chen, Yuting Wang, Qiankun Wang, Jingyi Shi, Huike Wang, Zichen Ye, Peng Xue, and Youlin Qiao. Impact of human and artificial intelligence collaboration on workload reduction in medical image interpretation. *npj Digital Medicine*, 7(1):349, 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01328-w. URL `https://doi.org/10.1038/s41746-024-01328-w`.

[80] Beomseok Sohn, K Y Park, J Choi, Josi H Koo, K Han, Bio Joo, Suk Yong Won, John Cha, H S Choi, and S.-K. Lee. Deep Learning–Based Software Improves Clinicians' Detection Sensitivity of Aneurysms on Brain TOF-MRA. *American Journal of Neuroradiology*, 42:1769 – 1775, 2021. URL `https://api.semanticscholar.org/CorpusID:236998019`.

[81] Mustaqueem Pallumeera, Jonathan C Giang, Ramanpreet Singh, Nooruddin S Pracha, and Mina S Makary. Evolving and Novel Applications of Artificial Intelligence in Cancer Imaging. *Cancers*, 17(9), 2025. ISSN 2072-6694. doi: 10.3390/cancers17091510. URL `https://www.mdpi.com/2072-6694/17/9/1510`.

[82] Luca Melazzini, Chandra Bortolotto, Leonardo Brizzi, Marina Achilli, Nicoletta Basla, Alessandro D'Onorio De Meo, Alessia Gerbasi, Olivia Maria Bottinelli, Riccardo Bellazzi, and Lorenzo Preda. AI for image quality and patient safety in CT and MRI. *European Radiology Experimental*, 9(1):28, 2025. ISSN 2509-9280. doi: 10.1186/s41747-025-00562-5. URL `https://doi.org/10.1186/s41747-025-00562-5`.

[83] Dr.Bhargavi Peddi Reddy, Doradla Bharadwaja, Mani Mohan Dupaty, Partha Sarkar, Dr. Mohammed Saleh, and Al Ansari. Using Generative Adversarial Networks and Ensemble Learning for Multi-Modal Medical Image Fusion to Improve the Diagnosis of Rare Neurological Disorders. *International Journal of Advanced Computer Science and Applications*, 2023. URL `https://api.semanticscholar.org/CorpusID:265580382`.

[84] Shenghan Zhang, Haoxuan Li, Ruixiang Tang, Sirui Ding, Laila Rasmy, Degui Zhi, Na Zou, and Xia Hu. PheME: A deep ensemble framework for improving phenotype prediction from multi-modal data. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pages 268–275, 2023. doi: 10.1109/ICHI57859.2023.00044.