The importance of domain-specific pre-training for reliable AI models

Abstract

Artificial Intelligence has been integrated into many sectors, which leads to accuracy and efficiency. To be precise, AI has improved reliability in multiple tasks like image classification, object detection, and segmentation where the automations of the systems can give a better output when compared to doing the same task manually. Healthcare is one of the most critical domains because even the slightest of the errors in any kind of diagnosis can highly impact the safety of the patient and the treatment planning and therefore, reliability is one of the most important factors for any AI system in the healthcare sector. In spite of knowing this, many AI models which are primarily trained on natural images are used in various healthcare applications making them even more dangerous in terms of reliability and accuracy. In this paper the importance of domain-specific pre-training data across four major imaging modalities is thoroughly examined, namely chest X-rays, magnetic resonance imaging (MRI), computer tomography (CT) and ultrasound. The AI models that are trained on domain-specific data show much better performance in terms of clinical decision-making and reliability. The findings demonstrate that domain-specific pre-training will not only benefit, but also be very essential for developing reliable and clinically useful AI systems in medical imaging and will also provide a clear direction for future research based on multi-modal data integration and real-world testing of AI in healthcare.

Keywords: AI, medical imaging, domain-specific pre-training, chest X-ray, MRI, CT, COVID-19, ultrasound, human-AI collaboration

Table 1: Overview of various Imaging Modalities

Modality	Key AI Tasks	Example Domain Specific Models	Challenges Addressed
Chest X-ray (CXR)	Pneumonia detection	CheXNet	Low resolution and overlapping structures
Computed Tomography (CT)	Lesion detection, cancer segmentation	3D U-Net, multi-view CNN	False positives, scanner variability
Magnetic Resonance Imaging (MRI)	Tumor segmentation, classification	nnU-Net, MRI-pretrained CNN	High soft tissue contrast and multi-sequence data
Ultrasound	Organ segmentation, echocardiography analysis	US-specific CNNs	Speckle noises and operator dependencies

Introduction

Medical imaging is vital for modern clinical diagnosis, allowing non-disruptive imaging of anatomical structures and any abnormal bodily function of cells/tissues, required for treatment, prognosis, and disease mapping. Imaging modalities such as chest X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Ultrasound are widely used in the healthcare industry and are one of the ultra-critical and compounded parts of that.

In recent years, Artificial Intelligence (AI), particularly deep learning (DL), has been a significant tool for medical image analysis, allowing noticeable improvements in diagnostic accuracy, efficiency, and clinical progress (Litjens et al., 2017). The success of deep learning in general computer vision has quickened the adoption of general-purpose foundational AI models in medical imaging, most of which are pre-trained on large-scale natural image datasets.

These AI models have shown exceptional accomplishment across a wide range of general visual tasks, which motivated their shift to medical imaging applications. As medical images are modality-specific, they greatly differ from natural images and thus require trained and expert clinical data interpretation. Consequently, AI models trained on the basis of non-medical data often fail to show effective results in the clinical field or practice. This raises concern regarding safety, accuracy, and dependability (Kelly et al., 2019; Roberts et al., 2021).

Every imaging modality faces challenges and difficulties and thus demands domain-aware learning. In the case of chest X-rays, there are two-dimensional representations with overlapping anatomical structures that require refined layered interpretation of distinct regions of the lungs, cardiac profile, and delicate radiographic patterns for the identification of diseases such as pneumonia and tuberculosis. Magnetic Resonance Imaging (MRI) offers a superior soft-tissue contrast through multiple phase sequences (e.g., T1, T2, T1ce), thereby enabling a comprehensive understanding of neurological and musculature abnormalities, thereby necessitating understanding of complicated tissue-specific magnitude variations.

General foundational AI models are still used in spite of such drawbacks. With the help of large-scale surveys, it has been found that general foundational AI models are often not accurate and have inadequate key features like clinical bases and lead to wrong explanations of anatomical features (Tajbakhsh et al., 2016; Litjens et al., 2017). This review emphasizes the need for domain specific pre-trained fundamentals for more reliable medical imaging AI models. With a proper combination of existing ideas and elements, this review paper equates general purpose foundational AI models on the basis of four major imaging modalities, namely- chest X-ray, MRI, CT and ultrasound.

Reliable automated analysis is difficult due to the multiplicative noise in ultrasound imaging, which involves the instant capturing, measuring, and processing of data. A comparison study clearly demonstrates the superiority of pre-trained domain-specific AI models. This paper not only demonstrates domain specific AI models as the bases of efficient human and AI collaboration in clinical practices, but also sets performance benchmarks. To improve the structure and the clarity of the paper, this review paper introduces a separate dedicated section on the physical and hardware principles of the various medical imaging modalities that is followed by AI focused sections in the following order: X-ray, CT, MRI and Ultrasound.

Hardware workings of the various medical imaging modalities and their limitations

The chest x-ray imaging is primarily based on the photons passing through the tissues of various densities present. While the thing is fast and cost effective, there are some limitations to it as well which is that the X-rays have many 2D projections with multiple overlapping anatomical structures which limit the depth perception and also the sensitivity factors to subtle pathologies such as early COVID-19 infiltrates and other kinds of severe lung diseases.

Computed tomography or CT, is widely used nowadays in the healthcare industry. The CT scans reconstruct multiple x-ray projections into volumetric 3-dimensional representations using various advanced demographic reconstruction algorithms and therefore it offers superior resolution and also shows a proper and a better density variation, but it also has a major limitation that is it exposes the patient to a higher amount of radiation doses which is extremely harmful.

Magnetic resonance imaging or MRI relies completely on the nuclear magnetic resonance principles to capture the tissue specific contrast through variations in the relaxation timings (T1, T2) and the pulse sequences. MRI also provides an exceptionally soft tissue contrast but on the other hand it suffers from extremely high cost, long scan times and also scanner dependent variability.

The last section, that is ultrasound imaging, uses high frequency sound waves to generate real time images. The advantage of this is that it is safe, portable and cheaper compared to the other imaging modalities but it is highly sensitive to speckle noises and operated dependency.

Chest X-Ray Imaging

One of the most widely used diagnostic modalities commonly used in clinical practices is the chest X-ray (CXR). This is generally used for the detection and also monitoring of pulmonary and cardiovascular diseases. Since the cost is low, the outcome is prompt, and it is also broadly available, chest radiology is considered to be the primary and the first imaging tool for detecting conditions such as pneumonia, tuberculosis, lung cancer, COVID-19, pleural effusion, heart failure and other such things.

However, chest X-rays are not free from complex analytical challenges. Plenty of such challenges are present for automated systems and AI models. One of the important reasons for this is that since the images are two-dimensional (2D), the overlapping anatomical structures and subtle intensity variations require expert interpretation. X-ray analysis is therefore very complex, and various studies show that the AI models pre-trained on domain-specific data most of the times outperform the general-purpose foundational vision AI models of comparable size especially in clinically sensitive tasks (Raghu et al., 2019; Litjens et al., 2017).

This type of domain-specific pre-training model becomes familiar with clinical terminologies, even the critical, advanced and complex ones. Not only this, but these models also become familiar with meaningful radiographic patterns, for example, lung-field asymmetries, interstitial markings and subtle opacities. These were generally missed or misinterpreted in the case of AI models that are pre-trained on natural images. General AI models that are pre-trained on natural images like ImageNet face problems detecting subtle signs of the disease since these models haven't seen any such medical images (Zech et al., 2018).

One of the most prominent examples of such training in the case of chest X-ray imaging is *CheXNet,* proposed by *Rajpurkar et. AI. (2017)*. Beyond pneumonia, various recent studies and literature have demonstrated the effectiveness of the usage of domain-specific AI models in detecting COVID-19 related things, tuberculosis cavitation and lung nodules as well, which are often subtle and easily missed by the general-purpose AI models. Since it was trained on large-scale expertly labelled X-ray datasets, CheXNet achieved state-of-the-art performance on multiple thoracic diseases, including pneumonia, with AUROC scores typically in the low-to-mid 0.80s on the ChestX-ray14 test set, comparable to radiologist-level performance (Rajpurkar et al., 2017).

However, AUROC alone is sometimes not sufficient for cross study comparison as the datasets, disease preventions differ greatly. It was also found that CheXNet consistently demonstrated highly superior diagnosis accuracy as compared to the general convolutional neural networks (CNNs). On the other hand, the performances of general-purpose CNNs pre-trained on natural images were low and exhibited poor generalization in clinical settings (Zech et al., 2018). Zech et al. (2018) also showed that the AI models that were trained on chest X-rays from a single institution were unable to generalize to external datasets. Especially when pretrained on non-medical images.

From these findings we can conclude that domain-specific pre-training will not only give more accurate results but will also reduce data bias to a large extent, therefore increasing robustness and reliability in real-world clinical environments. There are more recent studies that comparatively evaluated and found that the models pretrained on large-scale chest X-ray datasets consistently performed much better than general-purpose vision models of comparative size in both the classification and segmentation tasks (Irvin et al., 2019). The AI models that are pretrained on chest radiography are much aligned with radiological reasoning. Due to this the findings are better, and false positives are much less. This alignment is very important; since these are related to health, a small misclassification may lead to serious clinical consequences.

We can summaries the entire thing and say that extensive evidence has been confirmed in both foundational and recent studies that domain-specific pretraining is a significant factor on which AI performance in the case of chest X-ray imaging depends. These findings very clearly state that modality-aware pre-training is very important and essential for AI systems in chest radiography to be clinically trustworthy, and this provides a strong motivation to extend these domain-specific training strategies across other medical imaging modalities. On the large chest radiography data set there are 224,316 images of 65,240 patients. The deep learning models which were trained on chest X-ray images specifically achieved a higher multi pathology classification performance compared to the general Foundation AI models.

Table 2: Chest X-ray AI Model Comparison

Model	Pretraining	Task	Performance	Citation
CheXNet	Chest X-ray dataset	Pneumonia classification	High AUROC (≈0.82–0.85)	Rajpurkar et al., 2017
CNN (ImageNet)	ImageNet	Pneumonia classification	High AUROC (≈0.82–0.85)	Zech et al., 2018
CheXNet	Chest X-ray dataset	COVID-19, tuberculosis, nodules	Not explicitly quantified / Dataset-dependent	Litjens et al., 2017

Computed Tomography (CT) Imaging

Computed tomography imaging plays a very vital role in day-to-day clinical workflows, especially in oncology and trauma assessment tests. CT scan imaging is also widely used in the detection of pulmonary embolism, stroke tests, intracranial hemorrhage identification and COVID-19 tests as well. CT scans provide high-resolution three-dimensional (3D) volumetric data which capture very finely grained anatomical structures and details across a wide range of slices. These characteristics demand AI models which are capable of properly understanding the complex 3D spatial relationships, tissue density variations, and subtle pathological patterns.

Computed Tomography (CT) imaging represents one of the most complex modalities for reliable automated analysis. General-purpose foundational AI vision models which are trained on two-dimensional natural images, most of the time struggle to interpret the volumetric CT data effectively. These models lack intrinsic mechanisms to a great extent, which are necessary for learning spatial continuity across slices and thus very often fail to capture clinically relevant 3D contextual information.

Whereas in contrast, the AI models trained on volumetric CT data demonstrate highly improved sensitivity and robustness in the clinical detection and segmentation tasks (Çiçek et al., 2016; Setio et al., 2017). Previous studies demonstrate that CT-specific models which are trained on volumetric data improve cancer detection sensitivity and lesion localization to a great extent when compared to 2D vision models pretrained on natural images (Çiçek et al., 2016; Setio et al., 2017). Several studies show high improvements in sensitivity for lesion detection when CT-specific 3D architectures are used in comparison to 2D vision models (Setio et al., 2017).

CT-specific domain-trained models have been demonstrated to reduce false positives in lesion detection tasks, which is an extremely important and crucial factor in minimizing unnecessary follow-up procedures (Setio et al., 2017). These findings show the disadvantages of the general-purpose foundational AI models in handling volumetric medical data and also highlight the necessity of modality-aware pre-training.

Another major advantage of CT-specific domain pre-training is in its ability to account for scanner-related variability and reconstruction artefacts. The CT images vary a lot, depending a lot on the scanner manufacturers, the radiation dose protocols and the reconstruction algorithms. The pre-training of AI models with specific domain data enables the AI models to learn invariant features properly and in depth so that they can generalize across such variations, whereas general-purpose vision models often wrongly interpret reconstruction artefacts as pathological features and hence limit their clinical reliability (Litjens et al., 2017; Roberts et al., 2021).

Summing it up altogether, this research and these studies demonstrate that CT imaging shows fundamental shortcomings in general-purpose foundational vision AI models, and it also strongly supports the adoption of domain-specific data pretraining methods. For complex and critical applications like cancer detection and trauma diagnosis, the domain-aware CT AI models are extremely needed for achieving a clinically reliable and trustworthy AI performance.

Table 3: CT AI Performance

Model	Pre-training Type	Task	Performance	Citation
3D CNN	CT-specific dataset	Lesion detection	Dataset-dependent / High sensitivity reported	Setio et al., 2017
2D CNN	ImageNet	Lesion detection	Dataset-dependent / Low sensitivity reported	Setio et al., 2017

Magnetic Resonance Imaging (MRI)

Magnetic Resonance Imaging has emerged as one of the most important imaging tools in modern-day medical diagnosis, especially in the fields of neurological imaging, oncology, and musculoskeletal imaging. Unlike other medical imaging tools such as X-ray technology-based medical imaging modalities, the MRI image has very high-resolution volumetric data with very subtle contrast in soft tissues that are demonstrated through appropriate image acquisition protocols such as the T1-weighted image, the T2-weighted image, and the contrast-acquired image (T1ce image). This very multi-sequence and high-dimensional nature of the MRI data makes it in particular very unsuitable for direct transfer learning from the natural image datasets.

The analysis of an MRI image becomes more complex for a general AI system, this is because there is a large variation in the intensities specified to MRI image capturing tools which is again combined with a large amount of anatomical details which were absent in natural images. Due to this, MRI serves as a very important and crucial platform for evaluation of the effectiveness of domain-specific pre-training in medical imaging models. The transfer learning from the natural image datasets most of the times fail to capture the modality-specific intensity variations which are inherent to the MRIs.

A growing body of the existing literature continuously shows that the pre-trained AI models, those that are pre-trained on specific MRI datasets are much better than the general-purpose foundational AI models of same sizes across a wide range of tasks including tumor detection, classification and segmentation. Raghu et al. (2019) elaborately investigated the effectiveness of features learnt from the datasets which contained natural images in medical imaging tasks and demonstrated that the ImageNet pre-training provided very limited benefits for applications which were based on magnetic resonance imaging.

Their findings showed that when the AI models of the similar sizes were compared, the AI models which were pretrained directly on the MRI datasets consistently outperformed the ImageNet pretrained counterparts across medical tasks such as tumour detection and tumour segmentation. This limitation comes out most of the time because the MRI data contains a lot of modality-specific complex features and details, such as tissue contrast mechanisms and anatomical structures, which are definitely not present in the natural image datasets. Therefore, the general-purpose foundational AI models find it very difficult and complex to learn these complicated and meaningful MRI details and representations without proper domain exposures.

Studies support these findings by demonstrating that the AI models trained directly on MRI data achieve a greater and a better tumor segmentation performance across multiple sequences which include T1, T2, and also the contrast-enhanced MRI when compared to the AI models which are pretrained on natural images (Isensee et al., 2021; Bakas et al., 2018). This advantage in the performance has been continuously observed on the public brain tumour segmentation benchmarks such as BRATS.

Across the various public brain tumour segmentation benchmarks, MRI-specific training strategies, such as nnU-Net, have been reported to achieve Dice similarity coefficients around or above 0.80 on the BRATS benchmarks, depending on the task and validation split (Isensee et al., 2021), while the AI models relying mainly on ImageNet pretraining consistently show a lower segmentation accuracy (Raghu et al., 2019; Isensee et al., 2021). This throws light on the limitations of the ImageNet based pre-training for the MRI-specific segmentation tasks.

These differences in the performances are particularly significant in clinical matters, where proper and precise tumour boundary demarcation directly influences the treatment planning and the diagnostic outcomes. In addition to these segmentation tasks, the MRI-specific foundational AI models have also demonstrated a much better performance in classification tasks like tumour grading and the detection of diseases. These tasks include the stroke outcome prediction as well.

Multiple studies and literature have reported that the MRI-specific representational learning improves the disease classification and also the tumour grading performances when compared to the general-purpose foundational vision AI models, especially when the model sizes are the same (Raghu et al., 2019; Litjens et al., 2017). These improvements show the importance of AI models learning MRI-specific anatomical priors, which the general vision AI models often fail to capture effectively and properly. Other important and crucial factors influencing the MRI performances are scanner variability and acquisition heterogeneity. MRI images vary greatly across the scanner manufacturers, field strengths and patient demographics. Domain-specific pretraining with proper and wide MRI datasets exposes the AI models to the scanner and acquisition variability during learning, which leads to improved robustness and cross-institutional generalisation, whereas in contrast, the general-purpose models often fail to adapt reliably and effectively to the unseen MRI distributions (Raghu et al., 2019; Roberts et al., 2021). Such robustness is extremely crucial for the safety and the reliability factor of clinical deployment of the MRI based AI models.

Overall, the extensive empirical evidence confirms that the MRI-based tasks substantially benefit from domain-specific pretraining with MRI-specific data when the AI models of the similar sizes are compared. These findings properly highlight that the MRI's modality-specific complexity requires proper and appropriate learning strategies and that the reliable clinical deployment of AI in MRI cannot rely solely on the general-purpose vision pretraining.

Table 4: MRI AI Performance

Model	Pre-training Type	Task	Dice Score / AUROC	Citation
nnU-Net	MRI dataset	Brain tumor segmentation	High (≈0.80+)	Isensee et al., 2021
CNN	ImageNet	Brain tumor segmentation	Moderate (≈0.70–0.75)	Raghu et al., 2019

Ultrasound Imaging

Ultrasound imaging, due to its non-invasiveness, real-time acquisition and low cost, is being widely adopted across a wide range of clinical domains (e.g., abdominal imaging, obstetrics and cardiology). Other important applications include vascular assessment and echocardiography assessments as well. However, ultrasound imaging carries a number of complex and unique challenges, e.g., low signal-to-noise ratio, speckle noises, operator dependency, and considerable variability in image qualities. These characteristics make it extremely challenging for general foundational vision AI models trained on natural images.

Substantial empirical evidence in the literature has shown that general foundational vision AI models trained on general, natural images frequently mistake speckle noises and acquisition artefacts as meaningful features, resulting in unsafe and unreliable predictions. On the other hand, domain-specific pre-training on large-scale ultrasound data enables AI models to learn robust clinically relevant patterns despite the substantial noise and operator variability (Xie et al., 2018; Chen et al., 2019). Multiple surveys in the literature have shown that deep learning models pre-trained on domain-specific ultrasound data consistently yield strong performance compared to general foundational AI models with comparable parameter sizes (Xie et al., 2018; Chen et al., 2019).

The parameters such as accuracy, f1 score and robustness under the noise consistently favour the domain trained AI models. Specifically, in the noisy and operator-dependent imaging condition, AI models pre-trained on domain-specific pre-training data consistently outperform the general-purpose foundational AI models of comparable sizes (Xie et al., 2018; Chen et al., 2019).

These differences are very critical when AI is taken and applied under real-world clinical settings, especially when ultrasound images vary significantly across devices and patient populations. This significant performance gap between general-purpose foundational vision AI models and ultrasound-specific AI models accentuate the need for modality-aware training in AI models in the healthcare field for trustworthy, accurate, and reliable results.

Table 5: Ultrasound AI Performance

Model	Pre-training Type	Task	Ultrasound AI Performance	Citation
CNN	Ultrasound dataset	Abdominal/obstetric/cardiac	Not explicitly quantified / Dataset-dependent	Xie et al., 2018
CNN	ImageNet	Abdominal/obstetric/cardiac	Not explicitly quantified / Dataset-dependent	Xie et al., 2018

Discussion

Through the broad spectrum of the imaging modalities tested in this review paper, one could notice a similar trend stating that the AI models pre-trained with domain-specific medical AI data are clearly more powerful, reliable, and rationally sound when contrasted with foundational AI models. The medical images conclude that the modality specific details and pathologies are absent in the natural image datasets due to which they result in incorrect and unreliable results almost every time. Hence, we may conclude that the general-purpose foundational AI models are entirely relying on the mathematical explanations and the shortcuts in the datasets to a great extent in generalizing the practical applications in the healthcare sector (Tajbakhsh et al., 2016; Raghu et al., 2019; Zech et al., 2018; Oakden-Rayner, 2020; Roberts et al., 2021). Sometimes, the general vision models find it extremely difficult in terms of variations in image acquisition conditions, image reconstruction algorithms, and patient demographics as well. These variances are overcome with the help of domain-specific pre-trainings in AI models due to which more reliable interpretations are developed.

The domain-specific image datasets have also shown great future scope to improve. There are a few largely used image datasets in the domain-specific area which are not fully certified or verified in a correct manner by radiologists. Hence, a few annotations and residual variances are left behind (Rajpurkar et al., 2017). However, a few image datasets lack a variety of demographic populations due to which AI-related systems dedicated to health are causing alarming issues in terms of fairness and reliability (Oakden-Rayner, 2020; Roberts et al., 2021). To overcome these constraints, more interactions are needed between humans, doctors and AI researchers. Domain-specific AI models are not intended to replace doctors or radiologists. Instead, they serve as efficient tools to enhance human-AI collaboration. When these models are correctly developed with accurate training and used practically, they can act as a reliable second reader. This can help relieve radiologists from heavy workloads and aid in sound clinical decision-making. However, using radiologist-verified datasets is crucial for improving fairness, safety, and reliability.

The well-trained medical AI models can support radiologists by significantly reducing their workload and also acting as reliable and trustworthy readers. To understand the environments under which these AI models give clinically safe and trustworthy diagnostics results is extremely necessary for the application of AI in real world clinical practices.

Summing up this review, domain specific pre-training is not merely a decision-making technique, but also a basic requirement for the proper development of reliable and trustworthy AI models in the fields of medical practices especially medical imaging. By incorporating the findings of the various imaging modalities, this paper not only provides a comprehensive study of the existing literature, but also gives an insight for the future development process of strong, dependable and clinically channelized AI systems.

Conclusion

This review clearly demonstrates that the need for domain-specific pre-training is a significant necessity for the reliable AI system for medical images. The trained AI systems for modality-aware datasets have better standardized outcomes than the foundational vision AI systems for general applications. The outcomes also represent the significance of dependability for medical applications. The dependence on general vision AI systems provides a clearer understanding of the constraints of general vision AI systems for healthcare applications. The application of domain-specific Pre-trained AI systems would provide improved human-AI collaboration to align human behavior with better outcomes. Further research work would be required in multi-modal integration for successful utilization of AI systems for healthcare applications. Finally, the development of precise and reliable AI systems would be required for healthcare applications to incorporate understanding of the domain.

Acknowledgement

The author of this paper acknowledges the contributions of all the researchers whose work formed the foundation of this review paper and also extends his sincere gratitude to his mentors and peers for their valuable guidance and feedback all throughout this research process.

Bibliography

1. 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.
2. Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems, 32.
3. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225.
4. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine, 15(11), e1002683.
5. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019, July). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 590-597).
6. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods, 18(2), 203-211.
7. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., ... & Jambawalikar, S. R. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629.
8. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning?. IEEE transactions on medical imaging, 35(5), 1299-1312.
9. Setio, A. A. A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., Van Riel, S. J., ... & Van Ginneken, B. (2016). Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE transactions on medical imaging, 35(5), 1160-1169.
10. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016, October). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In International conference on medical image computing and computer-assisted intervention (pp. 424-432). Cham: Springer International Publishing.
11. Roth, H. R., Lu, L., Liu, J., Yao, J., Seff, A., Cherry, K., ... & Summers, R. M. (2015). Improving computer-aided detection using convolutional neural networks and random view aggregation. IEEE transactions on medical imaging, 35(5), 1170-1181.
12. Smistad, E., Falch, T. L., Bozorgi, M., Elster, A. C., & Lindseth, F. (2015). Medical image segmentation on GPUs–A comprehensive review. Medical image analysis, 20(1), 1-18.
13. Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S. X., ... & Wang, T. (2019). Deep learning in medical ultrasound analysis: a review. Engineering, 5(2), 261-275.

14. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC medicine, 17(1), 195.
15. Oakden-Rayner, L. (2020). Exploring large-scale public medical image datasets. Academic radiology, 27(1), 106-112.
16. Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., ... & Schönlieb, C. B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence, 3(3), 199-217.

The importance of domain-specific pre-training for reliable AI models, Author 100145

Overview: In this review, the author reviews the benefits of using pretrained AI models for medical imaging techniques such as MRI and CT imaging. The author correctly highlights the necessity for pretrained models to improve prediction accuracy. The author also does an excellent job of highlighting bias in training sets.

The manuscript could use rewriting at the level of grammatical mistakes, but also at the more macroscopic level of defining such important concepts as pretraining and eliminating repetition across the modalities discussed. The manuscript would benefit from a clearer discussion of how the methods work. I would accept after major revisions.

Substantive Comments:

1. I would recommend emphasizing the novelty of the ideas put forth in your article in the title. For example, I would argue that most practitioners already know that pretraining is important, but that's the focus of your title. Your novel angle should be reflected in the title. I would recommend highlighting the clinical aspect of your work in the title since I view that as the novelty.
2. I would move Table 1 into the text.
3. This study needs to be cited: “A comparison study clearly demonstrates the superiority of pre-trained domain-specific AI models.”
4. The manuscript would benefit from defining what pretraining is and going through more of the technical details before assuming familiarity on the subsequent modality discussions. I would recommend adding a section on ML and pre-training for medical imaging before going into the different modalities.
5. It would also benefit the article to specifically discuss why AI models fail without pre-training in each specific case and what pretraining can be done to improve them. Some technical details help.

Smaller Edits:

1. In the abstract, remove “To be precise” to increase flow.
2. Should be: “Healthcare is one of the most critical domains because even the slightest of the errors in any kind of diagnosis can highly impact the safety of the **patient and treatment planning; therefore,** reliability is one of the most important factors for any AI system in the healthcare sector.”
3. Delete “and are one of the ultra-critical and compounded parts of that.” Doesn't make sense. Also, one thing does use the plural verb are.
4. Should be challenges: “General foundational AI models are still used in spite of such challenges.”
5. The word equates does not make sense here: “With a proper combination of existing ideas and elements, this review paper equates general purpose foundational AI models on the basis of four major imaging modalities, namely- chest X-ray, MRI, CT and ultrasound.”
6. In many places you use the word “the”, where it shouldn't be used in English. Please use a grammar check to fix this.
7. Don't need the last section in “The last section, that is ultrasound imaging…”
8. Delete whereas in “Whereas in contrast”
9. Nothing was tested in this paper, everything was discussed: “Through the broad spectrum of the imaging modalities tested in this review paper”

Reviewer Report – Convergence Journal
Manuscript Title: *The Importance of Domain-Specific Pre-Training for Reliable AI Models*
Recommendation: Accept with Major Revisions

Manuscript Evaluation

This manuscript addresses a highly relevant and timely topic: the importance of domain-specific pre-training for improving the reliability of AI systems in medical imaging. The paper is well-motivated and clearly aligned with the journal's emphasis on rigor and empowerment, particularly as it seeks to guide responsible AI development in healthcare. The cross-modality structure (CXR, CT, MRI, ultrasound) is a strength and provides helpful comparative organization. The paper demonstrates good engagement with foundational literature (e.g., Raghu et al., Litjens et al., Roberts et al.) and appropriately highlights concerns about generalization, dataset bias, and clinical safety. The discussion of modality-specific challenges is thoughtful and pedagogically valuable. However, substantial revisions are required—particularly regarding research methods, evidentiary rigor, clarity of claims, and structural refinement—before the manuscript is ready for publication.

Major Strengths

1. Strong Conceptual Framing

The manuscript clearly articulates why medical imaging differs fundamentally from natural image domains and why this matters for transfer learning The modality-by-modality breakdown enhances clarity.

2. Broad Literature Engagement

The paper references influential works in transfer learning, domain generalization, and medical imaging benchmarks (e.g., CheXNet, nnU-Net, BRATS). This shows commendable effort in synthesizing the field.

3. Clinical Relevance

The discussion of real-world deployment challenges—scanner variability, demographic bias, annotation quality—is particularly strong and ethically grounded.

Major Revisions Needed (Research Methods Focus)

Although framed as a "review paper," the methodology for the review itself is not described. This is the most important area for improvement.

1. Clarify Review Methodology

Currently, the manuscript presents a narrative synthesis, but it does not specify:

- How were studies selected?
- What inclusion/exclusion criteria were used?

- Were only peer-reviewed articles included?
- Were performance metrics standardized for comparison?
- Was any systematic search strategy used?

Suggestion:
Add a dedicated “Methods” section explaining:

- Databases searched (e.g., PubMed, IEEE Xplore, arXiv)
- Keywords used
- Time frame of included studies
- Criteria for selecting comparative results
- Whether the review is systematic, scoping, or narrative

This will significantly strengthen the rigor and credibility of the manuscript.

2. Improve Quantitative Comparisons

Several tables report performance as “High,” “Moderate,” or “Dataset-dependent”. These descriptors are too vague for an academic review. Tables need to be within the test and have labelled and descriptions attached to them.

Suggestions:

- Provide actual reported metrics (AUROC, Dice, sensitivity, specificity).
- Clarify whether comparisons were made within the same dataset or across different datasets.
- Avoid implying superiority unless directly supported by controlled comparisons.

For example, Table 2 reports similar AUROC ranges for CheXNet and ImageNet CNNs (≈0.82–0.85) , yet the discussion claims superiority of domain-specific models. This inconsistency must be clarified.

3. Avoid Overgeneralization

The manuscript occasionally makes strong claims such as general models producing “incorrect and unreliable results almost every time”. These statements should be softened and supported with precise evidence.

Recommendation:

Rephrase to reflect nuance:

- Specify conditions where general models underperform.
- Distinguish between “limited transfer benefit” and “failure.”

Structural and Clarity Improvements

1. Reduce Repetition

Several arguments (e.g., modality-aware learning necessity) are repeated across sections. Consolidating key principles into a unified framework in the Discussion would improve flow.

2. Improve Writing Precision

There are frequent grammatical and stylistic issues:

- Redundant phrasing
- Informal constructions
- Long, repetitive sentences

While language clarity is acceptable overall, professional editing is recommended to elevate academic tone.

Minor Suggestions

- Clarify dataset statistics (e.g., 224,316 images of 65,240 patients; cite directly in text).
- Ensure consistency in citation formatting.
- Consider adding a short subsection on limitations of domain-specific pre-training (e.g., cost, data access barriers, regulatory constraints).

Final Recommendation: Accept with Major Revisions

This manuscript has strong potential and addresses an important and socially responsible topic. However, to meet high academic standards of rigor particularly regarding research methodology and evidentiary consistency substantial revision is necessary.

With clearer methodological transparency, improved quantitative precision, and refined argumentation, this paper could become a valuable contribution to discussions on trustworthy AI in healthcare.

I encourage the author to view this revision as an opportunity to further strengthen an already promising and meaningful piece of scholarship.

Enhancing the Clinical Reliability of AI models in Medical Imaging through the use of Domain-Specific Pre-Training Across Modalities

Abstract

Artificial Intelligence (AI) has been integrated into many sectors, which has led to improvements in accuracy and efficiency. AI has notably improved performances in various tasks such as image classification, object detection, and segmentation, most of the time outperforming manual approaches in different controlled environments. Healthcare is one of the most critical domains because even the slightest of errors in any kind of diagnosis can highly impact the safety of the patient and treatment planning therefore, reliability is one of the most important factors for any AI system in the healthcare sector. Despite this, many AI models, which are primarily trained on natural images, are used in various healthcare applications, which limits their clinical reliability and generalizability**.** In this paper, the importance of domain-specific pre-training data across four major imaging modalities is thoroughly examined, namely, Chest X-rays (CXR), Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and Ultrasound. The AI models that are trained on domain-specific data consistently show elevated performances in clinically applicable tasks, as reported across multiple studies**.** The findings demonstrate that domain-specific pre-training will not only benefit but will also be an indispensable element for developing reliable and clinically useful AI systems in medical imaging and will also provide a clear direction for future research based on multi-modal data integration and real-world testing of AI in healthcare.

Keywords: AI, Medical Imaging, Domain-Specific Pre-Training, Chest X-ray, MRI, CT, COVID-19, Ultrasound, Human-AI Collaboration

Introduction

Medical Imaging is vital for modern clinical diagnosis, allowing non-disruptive imaging of anatomical structures and any abnormal bodily function of cells/tissues, required for treatment, prognosis, and disease mapping. Imaging modalities such as Chest X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Ultrasound are widely used in the healthcare industry and are a critical component of the diagnostic workflow.

In recent years, Artificial Intelligence (AI), particularly deep learning (DL), has been a significant tool for medical image analysis, allowing noticeable improvements in diagnostic accuracy, efficiency, and clinical progress (Litjens et al., 2017). The success of deep learning in general computer vision has quickened the adoption of general-purpose foundational AI models in medical imaging, most of which are pre-trained on large-scale natural image datasets.

These AI models have shown exceptional accomplishment across a wide range of general visual tasks, which motivated their shift to medical imaging applications. As medical images are modality-specific, they greatly differ from natural images and thus require trained and expert clinical data interpretation. Consequently, AI models trained on the basis of non-medical data often fail to show effective results in the clinical field or practice. This raises concern regarding safety, accuracy, and dependability (Kelly et al., 2019; Roberts et al., 2021).

Every imaging modality faces challenges and difficulties and thus demands domain-aware learning. In the case of chest X-rays, there are two-dimensional representations with overlapping anatomical structures that require refined layered interpretation of distinct regions of the lungs, cardiac profile, and delicate radiographic patterns for the identification of diseases such as pneumonia and tuberculosis. Magnetic Resonance Imaging (MRI) offers a superior soft-tissue contrast through multiple phase sequences (e.g., T1, T2, T1ce), thereby enabling a comprehensive understanding of neurological and musculature abnormalities, necessitating an understanding of complicated tissue-specific magnitude variations.

General foundational AI models are still used in spite of such challenges. With the help of large-scale surveys, it has been found that general foundational AI models are often not accurate and have inadequate key features like clinical bases and lead to wrong explanations of anatomical features (Tajbakhsh et al., 2016; Litjens et al., 2017). This review emphasizes the need for domain specific pre-trained fundamentals for more reliable medical imaging AI models. With a proper

combination of existing ideas and elements, this review paper studies and analyzes general purpose foundational AI models on the basis of four major imaging modalities, namely: Chest X-ray, MRI, CT and Ultrasound.

A comparison across existing literature demonstrates the advantages of domain-specific pre-training in improving model robustness and clinical relevance (Litjens et al., 2017; Raghu et al., 2019).

Foundations of Pre-Training in Medical Imaging

The process of training an AI model on a large dataset to learn general feature representations before fine-tuning it on a specific task is referred to as “Pre-Training”. In medical imaging, general-purpose pre-training on foundational AI models (e.g. ImageNet) mostly provides an insubstantial benefit due to the fundamental differences between natural and medical images, which include modality specific intensity distributions and anatomical structures (Raghu et al., 2019).

Domain-specific pre-training in AI models addresses this limitation by training the models directly on suitable medical datasets, thereby enhancing the learning of clinically applicable features and improving robustness to modality-specific noise and variability. This study builds on the framework to evaluate how pre-training strategies guide model reliability across various imaging modalities.

Methodology

This study is a narrative review that synthesizes existing literature on the role of domain-specific pre-training of AI models in medical imaging. Relevant studies were identified through the different searches conducted on databases such as PubMed, IEEE Xplore, and arXiv using keywords which included “Medical Imaging”, “Transfer Learning”, “Domain-Specific Pre-Training”, “MRI Deep Learning”, “CT Segmentation”, “Ultrasound AI”, and “Chest X-ray Deep Learning”. The selection process prioritized peer-reviewed publications and widely cited research papers that provided empirical evaluation of AI models in medical imaging tasks such as image classification, detection, and object segmentation. Studies which compared domain-specific pre-training with general-purpose pre-training were included and discussed on the basis of modality-specific challenges in depth in this paper. Papers which lacked methodological clarity or empirical validation were excluded. This review is qualitative in nature and does not perform a formal meta-analysis; instead, performance metrics such as AUROC, Dice scores, sensitivity, and specificity are interpreted as mentioned in the original studies, with proper consideration of the dataset differences and experimental conditions.

Table 1: Overview of various Imaging Modalities

Modality	Key AI Tasks	Example Domain Specific Models	Challenges Addressed
Chest X-ray (CXR)	Pneumonia detection	CheXNet	Low resolution and overlapping structures
Computed Tomography (CT)	Lesion detection, cancer segmentation	3D U-Net, multi-view CNN	False positives, scanner variability
Magnetic Resonance Imaging (MRI)	Tumor segmentation, classification	nnU-Net, MRI-pretrained CNN	High soft tissue contrast and multi-sequence data
Ultrasound	Organ segmentation, echocardiography analysis	US-specific CNNs	Speckle noises and operator dependencies

Hardware workings of the various medical imaging modalities and their limitations

The chest x-ray imaging is primarily based on the photons passing through the tissues of various densities present. While this method is fast and cost-effective, there are some limitations to it as well, which is that the X-rays have many 2D projections with multiple overlapping anatomical structures, which limit the depth perception and also the sensitivity factors to subtle pathologies such as early COVID-19 infiltrates and other kinds of severe lung diseases.

Computed tomography or CT, is widely used currently in the healthcare industry. The CT scans reconstruct multiple X-ray projections into volumetric 3-dimensional representations using various advanced tomographic reconstruction algorithms and therefore it offers superior resolution and also shows a proper and a better density variation, but it also has a major limitation that is it exposes the patient to a higher amount of radiation doses which is extremely harmful.

Magnetic resonance imaging or MRI relies completely on the nuclear magnetic resonance principles to capture the tissue-specific contrast through variations in the relaxation timings (T1, T2) and the pulse sequences. MRI also provides an exceptionally soft tissue contrast but on the other hand it suffers from extremely high cost, long scan times and also scanner-dependent variability.

Ultrasound imaging, uses high frequency sound waves to generate real-time images. The advantage of this is that it is safe, portable and cheaper compared to the other imaging modalities but it is highly sensitive to speckle noises and operator dependency.

Chest X-Ray Imaging

One of the most widely used diagnostic modalities commonly used in clinical practices is the Chest X-ray (CXR). This is generally used for the detection and also monitoring of pulmonary and cardiovascular diseases. Since the cost is low, the outcome is prompt, and it is also broadly available, chest radiology is considered to be the primary and the first imaging tool for detecting conditions such as pneumonia, tuberculosis, lung cancer, COVID-19, pleural effusion, heart failure and other such things.

However, chest X-rays are not free from complex analytical challenges. Plenty of such challenges are present for automated systems and AI models. One of the important reasons for this is that since the images are two-dimensional (2D), the overlapping anatomical structures and subtle intensity variations require expert interpretation. X-ray analysis is therefore very complex, and various studies show that the AI models pre-trained on domain-specific data mostly surpass general-purpose foundational AI models in clinically relevant tasks when evaluated under comparable conditions (Raghu et al., 2019; Litjens et al., 2017).

This type of domain-specific pre-trained model becomes familiar with clinical terminologies, even the critical, advanced and complex ones. Not only this, but these models also become familiar with meaningful radiographic patterns, for example, lung-field asymmetries, interstitial markings and subtle opacities. These were generally missed or misinterpreted in the case of AI models that are pre-trained on natural images. General AI models that are pre-trained on natural images like ImageNet have difficulty perceiving subtle disease-specific patterns, due to negligible visibility to relevant medical imaging features (Zech et al., 2018).

One of the most prominent examples of such training in the case of chest X-ray imaging is *CheXNet,* proposed by *Rajpurkar et. al. (2017)*. Beyond pneumonia, various recent studies and literature have demonstrated the effectiveness of the usage of domain-specific AI models in detecting COVID-19 related things, tuberculosis cavitation and lung nodules as well, which are often subtle and easily missed by the general-purpose AI models. Since it was trained on large-scale expertly labelled X-ray datasets, CheXNet achieved state-of-the-art performance on multiple thoracic diseases, including pneumonia, with AUROC scores typically in the low-to-mid 0.80s on the ChestX-ray14 test set, comparable to radiologist-level performance (Rajpurkar et al., 2017).

However, AUROC alone is sometimes not sufficient for cross study comparison as the datasets, disease prevalence, and evaluation protocols differ greatly. It was also found that CheXNet illustrated improved diagnostic performances in several reviewed parameters as compared to the general convolutional neural networks (CNNs). On the other hand, the performances of general-purpose CNNs pre-trained on natural images were stated to show limited generalizability in certain clinical settings (Zech et al., 2018). Zech et al. (2018) also showed that the AI models that were trained on chest X-rays from a single institution were unable to generalize to external datasets. Especially when pretrained on non-medical images.

From these findings, we can conclude that domain-specific pre-training will not only give more accurate results but will also reduce data bias to a large extent, therefore increasing robustness and reliability in real-world clinical environments. There are more recent studies that comparatively evaluated and found that the models pretrained on large-scale chest X-ray datasets consistently performed much better than general-purpose vision models of comparative size in both the classification and segmentation tasks (Irvin et al., 2019). The AI models that are pretrained on chest radiography are much aligned with radiological reasoning. Due to this the findings are better, and false positives are much less. This alignment is very important; since these are related to health, a small misclassification may lead to serious clinical consequences.

These findings indicate that extensive evidence has been confirmed in both foundational and recent studies that domain-specific pre-training is a significant factor on which AI performance in the case of chest X-ray imaging depends. These findings very clearly state that modality-aware pre-training is very important and essential for AI systems in chest radiography to be clinically trustworthy, and this provides a strong motivation to extend these domain-specific training strategies across other medical imaging modalities. A large chest radiography dataset contains 224,316 images from 65,240 patients (Irvin et al., 2019). The deep learning models which were trained on chest X-ray images specifically achieved a higher multi-pathology classification performance compared to the general Foundation AI models.

Table 2: Chest X-ray AI Model Comparison

Model	Pre-training	Task	Performance	Notes
CheXNet	Chest X-ray dataset	Pneumonia classification	AUROC ≈ 0.82–0.85	Evaluated on ChestX-ray14 dataset
CNN	ImageNet	Pneumonia classification	Performance varies across datasets; often lower generalization in clinical settings	Generalization issues reported (Zech et al., 2018)
CheXNet	Chest X-ray dataset	Multi-disease detection	Performance depends on dataset and labeling quality	Supported by multiple studies

Computed Tomography (CT) Imaging

Computed tomography imaging plays a vital role in day-to-day clinical workflows, especially in oncology and trauma assessment tests. CT scan imaging is also widely used in the detection of pulmonary embolism, stroke tests, intracranial hemorrhage identification and COVID-19 tests as well. CT scans provide high-resolution three-dimensional (3D) volumetric data which capture very finely grained anatomical structures and details across a wide range of slices. These characteristics demand AI models which are capable of properly understanding the complex 3D spatial relationships, tissue density variations, and subtle pathological patterns.

Computed Tomography (CT) imaging represents one of the most complex modalities for reliable automated analysis. General-purpose foundational AI vision models which are trained on two-dimensional natural images, most of the time

struggle to interpret the volumetric CT data effectively. These models lack intrinsic mechanisms which are necessary for learning the spatial continuity across various slices and may sometimes fail to capture clinically relevant 3D contextual data.

In contrast, the AI models trained on volumetric CT data demonstrate highly improved sensitivity and robustness in the clinical detection and segmentation tasks (Çiçek et al., 2016; Setio et al., 2017). Previous studies demonstrate that CT-specific models which are trained on volumetric data improve cancer detection sensitivity and lesion localization to a great extent when compared to 2D vision models pretrained on natural images (Çiçek et al., 2016; Setio et al., 2017). Several studies show high improvements in sensitivity for lesion detection when CT-specific 3D architectures are used in comparison to 2D vision models (Setio et al., 2017).

CT-specific domain-trained models have been demonstrated to reduce false positives in lesion detection tasks, which is an extremely important and crucial factor in minimizing unnecessary follow-up procedures (Setio et al., 2017). These findings show the disadvantages of the general-purpose foundational AI models in handling volumetric medical data and also highlight the necessity of modality-aware pre-training.

Another major advantage of CT-specific domain pre-training is in its ability to account for scanner-related variability and reconstruction artefacts. The CT images vary a lot, depending a lot on the scanner manufacturers, the radiation dose protocols and the reconstruction algorithms. The pre-training of AI models with specific domain data enables the AI models to learn invariant features properly and in depth so that they can generalize across such variations, whereas general-purpose vision models often wrongly interpret reconstruction artefacts as pathological features and hence limit their clinical reliability (Litjens et al., 2017; Roberts et al., 2021).

Overall, this research and these studies demonstrate that CT imaging shows fundamental shortcomings in general-purpose foundational vision AI models, and it also strongly supports the adoption of domain-specific data pre-training methods. For complex and critical applications like cancer detection and trauma diagnosis, the domain-aware CT AI models are extremely needed for achieving a clinically reliable and trustworthy AI performance.

Table 3: CT AI Performance

Model	Pre-training Type	Task	Performance	Notes
3D CNN	CT-specific dataset	Lesion detection	Higher sensitivity reported in controlled studies	Captures volumetric context
2D CNN	ImageNet	Lesion detection	Lower sensitivity in volumetric tasks	Limited 3D understanding

Magnetic Resonance Imaging (MRI)

Magnetic Resonance Imaging has emerged as one of the most important imaging tools in modern-day medical diagnosis, especially in the fields of neurological imaging, oncology, and musculoskeletal imaging. Unlike other medical imaging tools such as X-ray technology-based medical imaging modalities, the MRI image has very high-resolution volumetric data with very subtle contrast in soft tissues that are demonstrated through appropriate image acquisition protocols such as the T1-weighted image, the T2-weighted image, and the contrast-acquired image (T1ce image). This very multi-sequence and high-dimensional nature of the MRI data makes it in particular less acceptable for direct transfer learning from the natural image datasets.

The analysis of an MRI image becomes more complex for a general AI system, this is because there is a large variation in the intensities specified to MRI image capturing tools which is again combined with a large amount of anatomical details which were absent in natural images. Due to this, MRI serves as a critical platform for evaluation of the effectiveness of

domain-specific pre-training in medical imaging models. The transfer learning from the natural image datasets most of the times provide a slight benefit in capturing the modality-specific intensity variations which are inherent to MRI data (Raghu et al., 2019).

A growing body of the existing literature continuously shows that the pre-trained AI models, those that are pre-trained on specific MRI datasets are much better than the general-purpose foundational AI models of same sizes across a wide range of tasks including tumor detection, classification and segmentation. Raghu et al. (2019) elaborately investigated the effectiveness of features learnt from the datasets which contained natural images in medical imaging tasks and demonstrated that the ImageNet pre-training provided very limited benefits for applications which were based on magnetic resonance imaging.

Their findings showed that when the AI models of the similar sizes were compared, the AI models which were pretrained directly on the MRI datasets consistently outperformed the ImageNet pretrained counterparts across medical tasks such as tumour detection and tumour segmentation. This limitation comes out most of the time because the MRI data contains a lot of modality-specific complex features and details, such as tissue contrast mechanisms and anatomical structures, which are definitely not present in the natural image datasets. Therefore, the general-purpose foundational AI models find it very difficult and complex to learn these complicated and meaningful MRI details and representations without proper domain exposures.

Studies support these findings by demonstrating that the AI models trained directly on MRI data achieve a greater and a better tumor segmentation performance across multiple sequences which include T1, T2, and also the contrast-enhanced MRI when compared to the AI models which are pretrained on natural images (Isensee et al., 2021; Bakas et al., 2018). This advantage in the performance has been continuously observed on the public brain tumor segmentation benchmarks such as BRATS.

Across the various public brain tumor segmentation benchmarks, MRI-specific training strategies, such as nnU-Net, have been reported to achieve Dice similarity coefficients around or above 0.80 on the BRATS benchmarks, depending on the task and validation split (Isensee et al., 2021), while the AI models relying mainly on ImageNet pre-training consistently show a lower segmentation accuracy (Raghu et al., 2019; Isensee et al., 2021). This throws light on the limitations of the ImageNet based pre-training for the MRI-specific segmentation tasks.

These differences in the performances are particularly significant in clinical matters, where proper and precise tumor boundary demarcation directly influences the treatment planning and the diagnostic outcomes. In addition to these segmentation tasks, the MRI-specific foundational AI models have also demonstrated a much better performance in classification tasks like tumour grading and the detection of diseases. These tasks include the stroke outcome prediction as well.

Multiple studies and literature have reported that the MRI-specific representational learning improves the disease classification and also the tumor grading performances when compared to the general-purpose foundational vision AI models, especially when the model sizes are the same (Raghu et al., 2019; Litjens et al., 2017). These improvements show the importance of AI models learning MRI-specific anatomical priors, which the general vision AI models often fail to capture effectively and properly. Other important and crucial factors influencing the MRI performances are scanner variability and acquisition heterogeneity. MRI images vary greatly across the scanner manufacturers, field strengths and patient demographics. Domain-specific pre-training with proper and wide MRI datasets exposes the AI models to the scanner and acquisition variability during learning, which leads to improved robustness and cross-institutional generalization, in contrast, the general-purpose models often fail to adapt reliably and effectively to the unseen MRI distributions (Raghu et al., 2019; Roberts et al., 2021). Such robustness is extremely crucial for the safety and the reliability factor of clinical deployment of the MRI based AI models.

Overall, the extensive empirical evidence confirms that the MRI-based tasks substantially benefit from domain-specific pre-training with MRI-specific data when the AI models of the similar sizes are compared. These findings properly highlight that the MRI's modality-specific complexity requires proper and appropriate learning strategies and that the reliable clinical deployment of AI in MRI cannot rely solely on the general-purpose vision pre-training.

Table 4: MRI AI Performance

Model	Pre-training Type	Task	Dice Score / AUROC	Notes
nnU-Net	MRI dataset	Brain tumor segmentation	≈ 0.80+	BRATS benchmark (Isensee et al., 2021)
CNN	ImageNet	Brain tumor segmentation	Typically lower than MRI-specific models	Limited transferability

Ultrasound Imaging

Ultrasound imaging, due to its non-invasiveness, real-time acquisition and low cost, is being widely adopted across a wide range of clinical domains (e.g., abdominal imaging, obstetrics and cardiology). Other important applications include vascular assessment and echocardiography assessments as well. However, ultrasound imaging carries a number of complex and unique challenges, e.g., low signal-to-noise ratio, speckle noises, operator dependency, and considerable variability in image qualities. The general foundational vision AI models which are trained on natural image datasets may misinterpret speckle noises and acquisition artefacts as meaningful features, thereby leading to unreliable predictions (Xie et al., 2018; Chen et al., 2019).

On the other hand, domain-specific pre-training on large-scale ultrasound data enables AI models to learn robust clinically relevant patterns despite the substantial noise and operator variability (Xie et al., 2018; Chen et al., 2019). Multiple surveys in the literature have shown that deep learning models pre-trained on domain-specific ultrasound data consistently yield strong performance compared to general foundational AI models with comparable parameter sizes (Xie et al., 2018; Chen et al., 2019).

The parameters such as accuracy, f1 score and robustness under the noise consistently favour the domain trained AI models. Specifically, in the noisy and operator-dependent imaging condition, AI models pre-trained on domain-specific pre-training data consistently outperform the general-purpose foundational AI models of comparable sizes (Xie et al., 2018; Chen et al., 2019).

These differences are very critical when AI is taken and applied under real-world clinical settings, especially when ultrasound images vary significantly across devices and patient populations. This significant performance gap between general-purpose foundational vision AI models and ultrasound-specific AI models accentuates the need for modality-aware training in AI models in the healthcare field for trustworthy, accurate, and reliable results.

Table 5: Ultrasound AI Performance

Model	Pre-training Type	Task	Performance	Notes
CNN	Ultrasound dataset	Clinical tasks	Improved robustness under noise conditions	Domain adaptation beneficial
CNN	ImageNet	Clinical tasks	Performance sensitive to noise and variability	Less robust

Discussion

Through the broad spectrum of the imaging modalities discussed in this review paper, one could notice a similar trend suggesting that AI models which are pre-trained on domain-specific medical data, often tend to be more powerful and reliable when contrasted with the foundational AI models. Various findings from medical imaging studies suggest that modality specific details and pathologies are most of the times absent in the natural image datasets due to which they may result in diminished performance in certain clinical tasks. Hence, we may conclude that the general-purpose foundational AI models rely on dataset-specific patterns and shortcuts to a great extent in generalizing the practical applications in the healthcare sector (Tajbakhsh et al., 2016; Raghu et al., 2019; Zech et al., 2018; Oakden-Rayner, 2020; Roberts et al., 2021). Sometimes, the general vision AI models may find it complex in terms of variations in image acquisition conditions, image reconstruction algorithms, and patient demographics as well. These variances are mostly better handled through domain-specific pre-trainings in AI models due to which more reliable interpretations are developed.

The domain-specific image datasets have also shown great future scope to improve. There are a few largely used image datasets in the domain-specific area which are not fully certified or verified in a correct manner by radiologists. Hence, residual variances may remain (Rajpurkar et al., 2017). However, a few image datasets lack a variety of demographic populations due to which AI-related systems dedicated to healthcare may raise concerns in terms of fairness and reliability (Oakden-Rayner, 2020; Roberts et al., 2021). To overcome these constraints, more interactions are needed between humans, doctors and AI researchers. Domain-specific AI models are not intended to replace doctors or radiologists. Instead, they serve as efficient tools to enhance human-AI collaboration. When these models are correctly developed with accurate training and used practically, they can act as a reliable second reader. This can help relieve radiologists from heavy workloads and aid in sound clinical decision-making. However, using radiologist-verified datasets is crucial for improving fairness, safety, and reliability.

The well-trained medical AI models can support radiologists by significantly reducing their workload and also acting as reliable readers. To understand the environments under which these AI models give clinically safe and trustworthy diagnostics results is extremely necessary for the application of AI in real world clinical practices.

Summing up this review, domain specific pre-training is not merely a decision-making technique, but also a basic requirement for the proper development of trustworthy AI models in the fields of medical practices especially medical imaging. By incorporating the findings of the various imaging modalities, this paper not only provides a comprehensive study of the existing literature, but also gives an insight for the future development process of strong, dependable and clinically channelized AI systems.

Limitations of Domain-Specific Pre-Training

Although domain-specific pre-training shows some promising advantages, some limitations have been introduced, which can be related to data availability, annotation costs, and legal issues. In addition, medical datasets need high-quality annotation, which is costly, and datasets may have limited demographic variation, which can affect fairness (Roberts et al., 2021). Finally, computational costs of large-scale domain-specific training are still high.

Conclusion

This review clearly demonstrates that the need for domain-specific pre-training is a significant necessity for the reliable AI system for medical images. The trained AI systems for modality-aware datasets have demonstrated a reliable and phenotypically accurate outputs than the foundational vision AI systems for general applications. The outcomes also represent the significance of dependability for medical applications. The dependence on general vision AI systems provides a clearer understanding of the constraints of general vision AI systems for healthcare applications. The application of domain-specific Pre-trained AI systems would provide improved human-AI collaboration to align human behavior with better outcomes. Further research work would be required in multi-modal integration for successful utilization of AI systems for

healthcare applications. Finally, the development of precise and reliable AI systems would be required for healthcare applications to incorporate understanding of the domain.

Acknowledgement

The author of this paper acknowledges the contributions of all the researchers whose work formed the foundation of this review paper and also extends his sincere gratitude to his mentors and peers for their valuable guidance and feedback all throughout this research process.

Bibliography

1. 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.
2. Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems, 32.
3. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225.
4. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine, 15(11), e1002683.
5. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019, July). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 590-597).
6. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods, 18(2), 203-211.
7. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., ... & Jambawalikar, S. R. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629.
8. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning?. IEEE transactions on medical imaging, 35(5), 1299-1312.
9. Setio, A. A. A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., Van Riel, S. J., ... & Van Ginneken, B. (2016). Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE transactions on medical imaging, 35(5), 1160-1169.
10. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016, October). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In International conference on medical image computing and computer-assisted intervention (pp. 424-432). Cham: Springer International Publishing.
11. Roth, H. R., Lu, L., Liu, J., Yao, J., Seff, A., Cherry, K., ... & Summers, R. M. (2015). Improving computer-aided detection using convolutional neural networks and random view aggregation. IEEE transactions on medical imaging, 35(5), 1170-1181.
12. Smistad, E., Falch, T. L., Bozorgi, M., Elster, A. C., & Lindseth, F. (2015). Medical image segmentation on GPUs–A comprehensive review. Medical image analysis, 20(1), 1-18.
13. Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S. X., ... & Wang, T. (2019). Deep learning in medical ultrasound analysis: a review. Engineering, 5(2), 261-275.
14. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC medicine, 17(1), 195.
15. Oakden-Rayner, L. (2020). Exploring large-scale public medical image datasets. Academic radiology, 27(1), 106-112.
16. Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., ... & Schönlieb, C. B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence, 3(3), 199-217.

Enhancing the Clinical Reliability of AI models in Medical Imaging through the use of Domain-Specific Pre-Training Across ModalitiesThe importance of domain-specific pre-training for reliable AI models¶

¶

Abstract

Artificial Intelligence (AI) has been integrated into many sectors, which has leads to improvements in accuracy and efficiency. To be precise, AI has notably improved reliability performances in multiple various tasks like such as image classification, object detection, and segmentation, most of the time outperforming manual approaches in different controlled environments. where the automations of the systems can give a better output when compared to doing the same task manually. Healthcare is one of the most critical domains because even the slightest of errors in any kind of diagnosis can highly impact the safety of the patient and treatment planning therefore, reliability is one of the most important factors for any AI system in the healthcare sector. Healthcare is one of the most critical domains because even the slightest of the errors in any kind of diagnosis can highly impact the safety of the patient and the treatment planning and therefore, reliability is one of the most important factors for any AI system in the healthcare sector. In spite of knowingDespite this, many AI models, which are primarily trained on natural images, are used in various healthcare applications, which limits their clinical reliability and generalizability. making them even more dangerous in terms of reliability and accuracy. In this paper, the importance of domain-specific pre-training data across four major imaging modalities is thoroughly examined,, namely, Cchest X-Rrays (CXR), Mmagnetic Rresonance Iimaging (MRI), Ccomputedr tTomography (CT) and Uultrasound. The AI models that are trained on domain-specific data consistently show much betterelevated performances in terms of clinical decision-making and reliabilityclinically applicable tasks, as reported across multiple studies. The findings demonstrate that domain-specific pre-training will not only benefit ,but will also be an indispensable element very essential for developing reliable and clinically useful AI systems in medical imaging and will also provide a clear direction for future research based on multi-modal data integration and real-world testing of AI in healthcare.

Keywords: AI, mMedical Iimaging, Ddomain-Sspecific Ppre-Ttraining, Cchest X-ray, MRI, CT, COVID-19, Uultrasound, hHuman-AI cCollaboration¶

¶

Table 1: Overview of various Imaging Modalities

Key AI Tasks¶	Example Domain-Specific Models¶	Challenges Addressed¶
Pneumonia detection¶	CheXNet¶	Low resolution and overlapping structures¶
Lesion detection, cancer segmentation¶	3D U-Net, multi-view CNN¶	False positives, scanner variability¶
Tumor segmentation, classification¶	nnU-Net, MRI-pretrained CNN¶	High soft tissue contrast and multi-sequence data¶
Organ segmentation, echocardiography analysis¶	US-specific CNNs¶	Speckle noises and operator dependencies¶

¶

Introduction

Medical imaging is vital for modern clinical diagnosis, allowing non-disruptive imaging of anatomical structures and any abnormal bodily function of cells/tissues, required for treatment, prognosis, and disease mapping. Imaging modalities such as Chest X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Ultrasound are widely used in the healthcare industry and are a critical component of the diagnostic workflow~~are one of the ultra-critical and compounded parts of that~~.

In recent years, Artificial Intelligence (AI), particularly deep learning (DL), has been a significant tool for medical image analysis, allowing noticeable improvements in diagnostic accuracy, efficiency, and clinical progress (Litjens et al., 2017). The success of deep learning in general computer vision has quickened the adoption of general-purpose foundational AI models in medical imaging, most of which are pre-trained on large-scale natural image datasets.

These AI models have shown exceptional accomplishment across a wide range of general visual tasks, which motivated their shift to medical imaging applications. As medical images are modality-specific, they greatly differ from natural images and thus require trained and expert clinical data interpretation. Consequently, AI models trained on the basis of non-medical data often fail to show effective results in the clinical field or practice. This raises concern regarding safety, accuracy, and dependability (Kelly et al., 2019; Roberts et al., 2021).

Every imaging modality faces challenges and difficulties and thus demands domain-aware learning. In the case of chest X-rays, there are two-dimensional representations with overlapping anatomical structures that require refined layered interpretation of distinct regions of the lungs, cardiac profile, and delicate radiographic patterns for the identification of diseases such as pneumonia and tuberculosis. Magnetic Resonance Imaging (MRI) offers a superior soft-tissue contrast through multiple phase sequences (e.g., T1, T2, T1ce), thereby enabling a comprehensive understanding of neurological and musculature abnormalities, ~~thereby~~ necessitating an understanding of complicated tissue-specific magnitude variations.

General foundational AI models are still used in spite of such ~~drawbacks~~challenges. With the help of large-scale surveys, it has been found that general foundational AI models are often not accurate and have inadequate key features like clinical bases and lead to wrong explanations of anatomical features (Tajbakhsh et al., 2016; Litjens et al., 2017). This review emphasizes the need for domain specific pre-trained fundamentals for more reliable medical imaging AI models. With a proper combination of existing ideas and elements, this review paper ~~equates~~ studies and analyzes general purpose foundational AI models on the basis of four major imaging modalities, namely: Chest X-ray, MRI, CT and Ultrasound.

A comparison across existing literature demonstrates the advantages of domain-specific pre-training in improving model robustness and clinical relevance (Litjens et al., 2017; Raghu et al., 2019).

Foundations of Pre-Training in Medical Imaging

The process of training an AI model on a large dataset to learn general feature representations before fine-tuning it on a specific task is referred to as “Pre-Training”. In medical imaging, general-purpose pre-training on foundational AI models (e.g. ImageNet) mostly provides an insubstantial benefit due to the fundamental differences between natural and medical images, which include modality specific intensity distributions and anatomical structures (Raghu et al., 2019).

Domain-specific pre-training in AI models addresses this limitation by training the models directly on suitable medical datasets, thereby enhancing the learning of clinically applicable features and improving robustness to modality-specific noise and variability. This study builds on the framework to evaluate how pre-training strategies guide model reliability across various imaging modalities.

Methodology

This study is a narrative review that synthesizes existing literature on the role of domain-specific pre-training of AI models in medical imaging. Relevant studies were identified through the different searches conducted on databases such as PubMed, IEEE Xplore, and arXiv using keywords which included “Medical Imaging”, “Transfer Learning”, “Domain-Specific Pre-Training”, “MRI Deep Learning”, “CT Segmentation”, “Ultrasound AI”, and “Chest X-ray Deep Learning”. The selection process prioritized peer-reviewed publications and widely cited research papers that provided empirical evaluation of AI models in medical imaging tasks such as image classification, detection, and object segmentation. Studies which compared domain-specific pre-training with general-purpose pre-training were included and discussed on the basis of modality-specific challenges in depth in this paper. Papers which lacked methodological clarity or empirical validation were excluded. This review is qualitative in nature and does not perform a formal meta-analysis; instead, performance metrics such as AUROC, Dice scores, sensitivity, and specificity are interpreted as mentioned in the original studies, with proper consideration of the dataset differences and experimental conditions.

~~Reliable automated analysis is difficult due to the multiplicative noise in ultrasound imaging, which involves the instant capturing, measuring, and processing of data. A comparison study clearly demonstrates the superiority of pre-trained domain-specific AI models. This paper not only demonstrates domain specific AI models as the bases of efficient human and AI collaboration in clinical practices, but also sets performance benchmarks. To improve the structure and the clarity of the paper, this review paper introduces a separate dedicated section on the physical and hardware principles of the various medical imaging modalities that is followed by AI focused sections in the following order: X-ray, CT, MRI and Ultrasound.~~¶

¶

Table 1: Overview of various Imaging Modalities

Modality	Key AI Tasks	Example Domain Specific Models	Challenges Addressed
Chest X-ray (CXR)	Pneumonia detection	CheXNet	Low resolution and overlapping structures
Computed Tomography (CT)	Lesion detection, cancer segmentation	3D U-Net, multi-view CNN	False positives, scanner variability
Magnetic Resonance Imaging (MRI)	Tumor segmentation, classification	nnU-Net, MRI-pretrained CNN	High soft tissue contrast and multi-sequence data
Ultrasound	Organ segmentation, echocardiography analysis	US-specific CNNs	Speckle noises and operator dependencies

¶

Hardware workings of the various medical imaging modalities and their limitations

The chest x-ray imaging is primarily based on the photons passing through the tissues of various densities present. While ~~the thing~~this method is fast and cost effective, there are some limitations to it as well, which is that the X-rays have many 2D projections with multiple overlapping anatomical structures, which limit the depth perception and also the sensitivity factors to subtle pathologies such as early COVID-19 infiltrates and other kinds of severe lung diseases.

Computed tomography or CT, is widely used ~~nowadays~~ currently in the healthcare industry. The CT scans reconstruct multiple X~~x~~-ray projections into volumetric 3-dimensional representations using various advanced ~~demographic~~ tomographic reconstruction algorithms and therefore it offers superior resolution and also shows a proper and a better density variation, but it also has a major limitation that is it exposes the patient to a higher amount of radiation doses which is extremely harmful.

Magnetic resonance imaging or MRI relies completely on the nuclear magnetic resonance principles to capture the tissue-specific contrast through variations in the relaxation timings (T1, T2) and the pulse sequences. MRI also provides an exceptionally soft tissue contrast but on the other hand it suffers from extremely high cost, long scan times and also scanner-dependent variability.

~~The last section, that is ultrasound~~Ultrasound imaging, uses high frequency sound waves to generate real-time images. The advantage of this is that it is safe, portable and cheaper compared to the other imaging modalities but it is highly sensitive to speckle noises and operator~~ed~~ dependency.

Chest X-Ray Imaging

One of the most widely used diagnostic modalities commonly used in clinical practices is the C~~c~~hest X-ray (CXR). This is generally used for the detection and also monitoring of pulmonary and cardiovascular diseases. Since the cost is low, the outcome is prompt, and it is also broadly available, chest radiology is considered to be the primary and the first imaging tool for detecting conditions such as pneumonia, tuberculosis, lung cancer, COVID-19, pleural effusion, heart failure and other such things.

However, chest X-rays are not free from complex analytical challenges. Plenty of such challenges are present for automated systems and AI models. One of the important reasons for this is that since the images are two-dimensional (2D), the overlapping anatomical structures and subtle intensity variations require expert interpretation. X-ray analysis is therefore very complex, and various studies show that the AI models pre-trained on domain-specific data most of the times outperform the general-purpose foundational vision AI models of comparable size especially in clinically sensitive tasks (Raghu et al., 2019; Litjens et al., 2017).

This type of domain-specific pre-training model becomes familiar with clinical terminologies, even the critical, advanced and complex ones. Not only this, but these models also become familiar with meaningful radiographic patterns, for example, lung-field asymmetries, interstitial markings and subtle opacities. These were generally missed or misinterpreted in the case of AI models that are pre-trained on natural images. General AI models that are pre-trained on natural images like ImageNet have difficulty perceiving subtle disease-specific patterns, due to negligible visibility to relevant medical imaging features (Zech et al., 2018). ~~face problems detecting subtle signs of the disease since these models haven't seen any such medical images (Zech et al., 2018).~~

One of the most prominent examples of such training in the case of chest X-ray imaging is *CheXNet,* proposed by *Rajpurkar et. Al. (2017)*. Beyond pneumonia, various recent studies and literature have demonstrated the effectiveness of the usage of domain-specific AI models in detecting COVID-19 related things, tuberculosis cavitation and lung nodules as well, which are often subtle and easily missed by the general-purpose AI models. Since it was trained on large-scale expertly labelled X-ray datasets, CheXNet achieved state-of-the-art performance on multiple thoracic diseases, including pneumonia, with AUROC scores typically in the low-to-mid 0.80s on the ChestX-ray14 test set, comparable to radiologist-level performance (Rajpurkar et al., 2017).

However, AUROC alone is sometimes not sufficient for cross study comparison as the datasets, disease ~~preventions~~ prevalence, and evaluation protocols differ greatly. It was also found that CheXNet ~~consistently~~ illustrated improved diagnostic performances in several reviewed parameter as compared to ~~demonstrated highly superior diagnosis accuracy as compared to~~ the general convolutional neural networks (CNNs). On the other hand, the performances of general-purpose CNNs pre-trained on natural images were stated to show limited generalizability in certain clinical settings ~~low and exhibited poor generalization in clinical settings~~ (Zech et al., 2018). Zech et al. (2018) also showed that

the AI models that were trained on chest X-rays from a single institution were unable to generalize to external datasets. Especially when pretrained on non-medical images.

From these findings, we can conclude that domain-specific pre-training will not only give more accurate results but will also reduce data bias to a large extent, therefore increasing robustness and reliability in real-world clinical environments. There are more recent studies that comparatively evaluated and found that the models pretrained on large-scale chest X-ray datasets consistently performed much better than general-purpose vision models of comparative size in both the classification and segmentation tasks (Irvin et al., 2019). The AI models that are pretrained on chest radiography are much aligned with radiological reasoning. Due to this the findings are better, and false positives are much less. This alignment is very important; since these are related to health, a small misclassification may lead to serious clinical consequences.

These findings indicate that ~~We can summaries the entire thing and say that~~ extensive evidence has been confirmed in both foundational and recent studies that domain-specific pretraining is a significant factor on which AI performance in the case of chest X-ray imaging depends. These findings very clearly state that modality-aware pre-training is very important and essential for AI systems in chest radiography to be clinically trustworthy, and this provides a strong motivation to extend these domain-specific training strategies across other medical imaging modalities. ~~On the~~A large chest radiography data set there are 224,316 images of 65,240 patients (Irvin et al., 2019).~~:~~ ~~-~~The deep learning models which were trained on chest X-ray images specifically achieved a higher multi~~-~~pathology classification performance compared to the general Foundation AI models.

Table 2: Chest X-ray AI Model Comparison

Model	Pretraining	Task	Performance	~~Citation~~Notes
CheXNet	Chest X-ray dataset	Pneumonia classification	AUROC ≈ 0.82–0.85~~High AUROC (≈0.82–0.85)~~	Evaluated on ChestX-ray14 dataset~~Rajpurkar et al., 2017~~
CNN ~~(ImageNet)~~	ImageNet	Pneumonia classification	Performance varies across datasets; often lower generalization in clinical settings~~High AUROC (≈0.82–0.85)~~	Generalization issues reported (Zech et al., 2018)~~Zech et al., 2018~~
CheXNet	Chest X-ray dataset	Multi-disease detection ~~COVID-19, tuberculosis, nodules~~	Performance depends on dataset and labeling quality~~Not explicitly quantified / Dataset-dependent~~	Supported by multiple studies~~Litjens et al., 2017~~

Computed Tomography (CT) Imaging

Computed tomography imaging plays a very vital role in day-to-day clinical workflows, especially in oncology and trauma assessment tests. CT scan imaging is also widely used in the detection of pulmonary embolism, stroke tests, intracranial hemorrhage identification and COVID-19 tests as well. CT scans provide high-resolution three-dimensional (3D) volumetric data which capture very finely grained anatomical structures and details across a wide range of slices. These characteristics demand AI models which are capable of properly understanding the complex 3D spatial relationships, tissue density variations, and subtle pathological patterns.

Computed Tomography (CT) imaging represents one of the most complex modalities for reliable automated analysis. General-purpose foundational AI vision models which are trained on two-dimensional natural images, most of the time struggle to interpret the volumetric CT data effectively. These models lack intrinsic mechanisms which are ~~to a great extent, which are~~ necessary for learning spatial continuity across various slices and may ~~thus very often~~ sometimes fail to capture clinically relevant 3D contextual ~~information~~data.

~~Whereas i~~In contrast, the AI models trained on volumetric CT data demonstrate highly improved sensitivity and robustness in the clinical detection and segmentation tasks (Çiçek et al., 2016; Setio et al., 2017). Previous studies demonstrate that CT-specific models which are trained on volumetric data improve cancer detection sensitivity and lesion localization to a great extent when compared to 2D vision models pretrained on natural images (Çiçek et al., 2016; Setio et al., 2017). Several studies show high improvements in sensitivity for lesion detection when CT-specific 3D architectures are used in comparison to 2D vision models (Setio et al., 2017).

CT-specific domain-trained models have been demonstrated to reduce false positives in lesion detection tasks, which is an extremely important and crucial factor in minimizing unnecessary follow-up procedures (Setio et al., 2017). These findings show the disadvantages of the general-purpose foundational AI models in handling volumetric medical data and also highlight the necessity of modality-aware pre-training.

Another major advantage of CT-specific domain pre-training is in its ability to account for scanner-related variability and reconstruction artefacts. The CT images vary a lot, depending a lot on the scanner manufacturers, the radiation dose protocols and the reconstruction algorithms. The pre-training of AI models with specific domain data enables the AI models to learn invariant features properly and in depth so that they can generalize across such variations, whereas general-purpose vision models often wrongly interpret reconstruction artefacts as pathological features and hence limit their clinical reliability (Litjens et al., 2017; Roberts et al., 2021).

~~Summing it up altogether~~Overall, this research and these studies demonstrate that CT imaging shows fundamental shortcomings in general-purpose foundational vision AI models, and it also strongly supports the adoption of domain-specific data pretraining methods. For complex and critical applications like cancer detection and trauma diagnosis, the domain-aware CT AI models are extremely needed for achieving a clinically reliable and trustworthy AI performance.

Table 3: CT AI Performance

Model	Pre-training Type	Task	Performance	Notes~~Citation~~
3D CNN	CT-specific dataset	Lesion detection	Higher sensitivity reported in controlled studies~~Dataset-dependent / High sensitivity reported~~	Captures volumetric context~~Setio et al., 2017~~
2D CNN	ImageNet	Lesion detection	Lower sensitivity in volumetric tasks~~Dataset-dependent / Low sensitivity reported~~	Limited 3D understanding~~Setio et al., 2017~~

Magnetic Resonance Imaging (MRI)

Magnetic Resonance Imaging has emerged as one of the most important imaging tools in modern-day medical diagnosis, especially in the fields of neurological imaging, oncology, and musculoskeletal imaging. Unlike other medical imaging tools such as X-ray technology-based medical imaging modalities, the MRI image has very high-resolution volumetric data with

very subtle contrast in soft tissues that are demonstrated through appropriate image acquisition protocols such as the T1-weighted image, the T2-weighted image, and the contrast-acquired image (T1ce image). This very multi-sequence and high-dimensional nature of the MRI data makes it in particular very unsuitable for direct transfer learning from the natural image datasets.

The analysis of an MRI image becomes more complex for a general AI system, this is because there is a large variation in the intensities specified to MRI image capturing tools which is again combined with a large amount of anatomical details which were absent in natural images. Due to this, MRI serves as a ~~very important and crucial platform~~critical platform for evaluation of the effectiveness of domain-specific pre-training in medical imaging models. The transfer learning from the natural image datasets most of the times provide a slight benefit in capturing the modality-specific intensity variations which are inherent to MRI data (Raghu et al., 2019).~~fail to capture the modality-specific intensity variations which are inherent to the MRIs.~~

A growing body of the existing literature continuously shows that the pre-trained AI models, those that are pre-trained on specific MRI datasets are much better than the general-purpose foundational AI models of same sizes across a wide range of tasks including tumor detection, classification and segmentation. Raghu et al. (2019) elaborately investigated the effectiveness of features learnt from the datasets which contained natural images in medical imaging tasks and demonstrated that the ImageNet pre-training provided very limited benefits for applications which were based on magnetic resonance imaging.

Their findings showed that when the AI models of the similar sizes were compared, the AI models which were pretrained directly on the MRI datasets consistently outperformed the ImageNet pretrained counterparts across medical tasks such as tumour detection and tumour segmentation. This limitation comes out most of the time because the MRI data contains a lot of modality-specific complex features and details, such as tissue contrast mechanisms and anatomical structures, which are definitely not present in the natural image datasets. Therefore, the general-purpose foundational AI models find it very difficult and complex to learn these complicated and meaningful MRI details and representations without proper domain exposures.

Studies support these findings by demonstrating that the AI models trained directly on MRI data achieve a greater and a better tumor segmentation performance across multiple sequences which include T1, T2, and also the contrast-enhanced MRI when compared to the AI models which are pretrained on natural images (Isensee et al., 2021; Bakas et al., 2018). This advantage in the performance has been continuously observed on the public brain tumour segmentation benchmarks such as BRATS.

Across the various public brain tumour segmentation benchmarks, MRI-specific training strategies, such as nnU-Net, have been reported to achieve Dice similarity coefficients around or above 0.80 on the BRATS benchmarks, depending on the task and validation split (Isensee et al., 2021), while the AI models relying mainly on ImageNet pre-training consistently show a lower segmentation accuracy (Raghu et al., 2019; Isensee et al., 2021). ~~This throws~~This throws light on the limitations of the ImageNet based pre-training for the MRI-specific segmentation tasks.

These differences in the performances are particularly significant in clinical matters, where proper and precise ~~tumour~~tumor boundary demarcation directly influences the treatment planning and the diagnostic outcomes. In addition to these segmentation tasks, the MRI-specific foundational AI models have also demonstrated a much better performance in classification tasks like ~~tumour~~tumor grading and the detection of diseases. These tasks include the stroke outcome prediction as well.

Multiple studies and literature have reported that the MRI-specific representational learning improves the disease classification and also the ~~tumour~~tumor grading performances when compared to the general-purpose foundational vision AI models, especially when the model sizes are the same (Raghu et al., 2019; Litjens et al., 2017). These improvements show the importance of AI models learning MRI-specific anatomical priors, which the general vision AI models often fail to capture effectively and properly. Other important and crucial factors influencing the MRI performances are scanner variability and acquisition heterogeneity. MRI images vary greatly across the scanner manufacturers, field strengths and patient demographics. Domain-specific pretraining with proper and wide MRI datasets exposes the AI models to the scanner and acquisition variability during learning, which leads to improved robustness and cross-institutional

generalisationgeneralization, whereas in contrast, the general-purpose models often fail to adapt reliably and effectively to the unseen MRI distributions (Raghu et al., 2019; Roberts et al., 2021). Such robustness is extremely crucial for the safety and the reliability factor of clinical deployment of the MRI based AI models.

Overall, the extensive empirical evidence confirms that the MRI-based tasks substantially benefit from domain-specific pre-training with MRI-specific data when the AI models of the similar sizes are compared. These findings properly highlight that the MRI's modality-specific complexity requires proper and appropriate learning strategies and that the reliable clinical deployment of AI in MRI cannot rely solely on the general-purpose vision pretraining.¶

Table 4: MRI AI Performance

Model	Pre-training Type	Task	Dice Score / AUROC	CitationNotes
nnU-Net	MRI dataset	Brain tumor segmentation	≈ 0.80+High (≈0.80+)	BRATS benchmark (Isensee et al., 2021)Isensee et al., 2021
CNN	ImageNet	Brain tumor segmentation	Typically lower than MRI-specific modelsModerate (≈0.70–0.75)	Limited transferabilityRaghu et al., 2019

Ultrasound Imaging

Ultrasound imaging, due to its non-invasiveness, real-time acquisition and low cost, is being widely adopted across a wide range of clinical domains (e.g., abdominal imaging, obstetrics and cardiology). Other important applications include vascular assessment and echocardiography assessments as well. However, ultrasound imaging carries a number of complex and unique challenges, e.g., low signal-to-noise ratio, speckle noises, operator dependency, and considerable variability in image qualities. The general foundational vision AI models which are trained on natural image datasets may misinterpret speckle noises and acquisition artefacts as meaningful features, thereby leading to unreliable predictions (Xie et al., 2018; Chen et al., 2019).These characteristics make it extremely challenging for general foundational vision AI models trained on natural images.

On the other hand, domain-specific pre-training on large-scale ultrasound data enables AI models to learn robust clinically relevant patterns despite the substantial noise and operator variability (Xie et al., 2018; Chen et al., 2019). Multiple surveys in the literature have shown that deep learning models pre-trained on domain-specific ultrasound data consistently yield strong performance compared to general foundational AI models with comparable parameter sizes (Xie et al., 2018; Chen et al., 2019).

Substantial empirical evidence in the literature has shown that general foundational vision AI models trained on general, natural images frequently mistake speckle noises and acquisition artefacts as meaningful features, resulting in unsafe and unreliable predictions. On the other hand, domain-specific pre-training on large-scale ultrasound data enables AI models to learn robust clinically relevant patterns despite the substantial noise and operator variability (Xie et al., 2018; Chen et al., 2019). Multiple surveys in the literature have shown that deep learning models pre-trained on domain-specific ultrasound data consistently yield strong performance compared to general foundational AI models with comparable parameter sizes (Xie et al., 2018; Chen et al., 2019). ¶

The parameters such as accuracy, f1 score and robustness under the noise consistently favour the domain trained AI models. Specifically, in the noisy and operator-dependent imaging condition, AI models pre-trained on domain-specific

pre-training data consistently outperform the general-purpose foundational AI models of comparable sizes (Xie et al., 2018; Chen et al., 2019).

These differences are very critical when AI is taken and applied under real-world clinical settings, especially when ultrasound images vary significantly across devices and patient populations. This significant performance gap between general-purpose foundational vision AI models and ultrasound-specific AI models accentuate the need for modality-aware training in AI models in the healthcare field for trustworthy, accurate, and reliable results.

Table 5: Ultrasound AI Performance

Model	Pre-training Type	Task	~~Ultrasound AI~~ Performance	Notes~~Citation~~
CNN	Ultrasound dataset	Clinical tasks~~Abdominal/obstetric/cardiac~~	Improved robustness under noise conditions~~Not explicitly quantified / Dataset-dependent~~	Domain adaptation beneficial~~Xie et al., 2018~~
CNN	ImageNet	Clinical tasks~~Abdominal/obstetric/cardiac~~	Performance sensitive to noise and variability~~Not explicitly quantified / Dataset-dependent~~	Less robust~~Xie et al., 2018~~

Discussion

Through the broad spectrum of the imaging modalities tested in this review paper, one could notice a similar trend ~~stating~~ suggesting that ~~the~~ AI models which are pre-trained ~~with~~ on domain-specific medical ~~AI~~ data, often tend to be more powerful and reliable ~~are clearly more powerful, reliable, and rationally sound~~ when contrasted with the foundational AI models. Various findings from medical imaging studies suggest that modality specific details and pathologies are most of the times absent in the natural image datasets due to which they may result in diminished performance in certain clinical tasks. ~~The medical images conclude that the modality specific details and pathologies are absent in the natural image datasets due to which they result in incorrect and unreliable results almost every time.~~ Hence, we may conclude that the general-purpose foundational AI models ~~are entirely relying~~rely on dataset-specific patterns and shortcuts~~the mathematical explanations and the shortcuts in the datasets~~ to a great extent in generalizing the practical applications in the healthcare sector (Tajbakhsh et al., 2016; Raghu et al., 2019; Zech et al., 2018; Oakden-Rayner, 2020; Roberts et al., 2021). Sometimes, the general vision models find it extremely difficult in terms of variations in image acquisition conditions, image reconstruction algorithms, and patient demographics as well. These variances are ~~overcome~~ mostly better handled~~with the help of~~ through domain-specific pre-trainings in AI models due to which more reliable interpretations are developed.

The domain-specific image datasets have also shown great future scope to improve. There are a few largely used image datasets in the domain-specific area which are not fully certified or verified in a correct manner by radiologists. Hence, ~~a few annotations and res~~residual variances ~~are left behind~~may remain (Rajpurkar et al., 2017). However, a few image datasets lack a variety of demographic populations due to which AI-related systems dedicated to healthcare may raise concerns in terms of fairness and ~~are causing alarming issues in terms of fairness and reliability~~ reliability (Oakden-Rayner, 2020; Roberts et al., 2021). To overcome these constraints, more interactions are needed between humans, doctors and AI researchers. Domain-specific AI models are not intended to replace doctors or radiologists. Instead, they serve as efficient tools to enhance human-AI collaboration. When these models are correctly developed with accurate training and used practically, they can act as a reliable second reader. This can help relieve radiologists from

heavy workloads and aid in sound clinical decision-making. However, using radiologist-verified datasets is crucial for improving fairness, safety, and reliability.

The well-trained medical AI models can support radiologists by significantly reducing their workload and also acting as reliable and trustworthy readers. To understand the environments under which these AI models give clinically safe and trustworthy diagnostics results is extremely necessary for the application of AI in real world clinical practices.

Summing up this review, domain specific pre-training is not merely a decision-making technique, but also a basic requirement for the proper development of ~~reliable and~~ trustworthy AI models in the fields of medical practices especially medical imaging. By incorporating the findings of the various imaging modalities, this paper not only provides a comprehensive study of the existing literature, but also gives an insight for the future development process of strong, dependable and clinically channelized AI systems.

Limitations of Domain-Specific Pre-Training

Although domain-specific pre-training shows some promising advantages, some limitations have been introduced, which can be related to data availability, annotation costs, and legal issues. In addition, medical datasets need high-quality annotation, which is costly, and datasets may have limited demographic variation, which can affect fairness (Roberts et al., 2021). Finally, computational costs of large-scale domain-specific training are still high.

Conclusion

This review clearly demonstrates that the need for domain-specific pre-training is a significant necessity for the reliable AI system for medical images. The trained AI systems for modality-aware datasets have demonstrated a reliable and phenotypically accurate outputs~~better standardized outcomes~~ than the foundational vision AI systems for general applications. The outcomes also represent the significance of dependability for medical applications. The dependence on general vision AI systems provides a clearer understanding of the constraints of general vision AI systems for healthcare applications. The application of domain-specific Pre-trained AI systems would provide improved human-AI collaboration to align human behavior with better outcomes. Further research work would be required in multi-modal integration for successful utilization of AI systems for healthcare applications. Finally, the development of precise and reliable AI systems would be required for healthcare applications to incorporate understanding of the domain.

Acknowledgement

The author of this paper acknowledges the contributions of all the researchers whose work formed the foundation of this review paper and also extends his sincere gratitude to his mentors and peers for their valuable guidance and feedback all throughout this research process.

Bibliography

1. 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.
2. Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems, 32.
3. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225.
4. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine, 15(11), e1002683.

5. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019, July). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 590-597).
6. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods, 18(2), 203-211.
7. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., ... & Jambawalikar, S. R. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629.
8. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning?. IEEE transactions on medical imaging, 35(5), 1299-1312.
9. Setio, A. A. A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., Van Riel, S. J., ... & Van Ginneken, B. (2016). Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE transactions on medical imaging, 35(5), 1160-1169.
10. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016, October). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In International conference on medical image computing and computer-assisted intervention (pp. 424-432). Cham: Springer International Publishing.
11. Roth, H. R., Lu, L., Liu, J., Yao, J., Seff, A., Cherry, K., ... & Summers, R. M. (2015). Improving computer-aided detection using convolutional neural networks and random view aggregation. IEEE transactions on medical imaging, 35(5), 1170-1181.
12. Smistad, E., Falch, T. L., Bozorgi, M., Elster, A. C., & Lindseth, F. (2015). Medical image segmentation on GPUs–A comprehensive review. Medical image analysis, 20(1), 1-18.
13. Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S. X., ... & Wang, T. (2019). Deep learning in medical ultrasound analysis: a review. Engineering, 5(2), 261-275.
14. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC medicine, 17(1), 195.
15. Oakden-Rayner, L. (2020). Exploring large-scale public medical image datasets. Academic radiology, 27(1), 106-112.
16. Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., ... & Schönlieb, C. B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence, 3(3), 199-217.

Response to Reviewer 1

I sincerely extend my heartful thanks and gratitude to the reviewer for providing me such a detailed and constructive feedback. I appreciate the positive evaluation of my work and recognising the importance of the domain-specific pre-training and above all the discussion on the bias in the training datasets. I have thoroughly revised the manuscript and addressed all the required changes. I have also tried to improve the quality of writing, structure and the clarity of the contents.

Substantive Comments

Comment 1: I would recommend emphasizing the novelty of the ideas put forth in your article in the title. For example, I would argue that most practitioners already know that pretraining is important, but that's the focus of your title. Your novel angle should be reflected in the title. I would recommend highlighting the clinical aspect of your work in the title since I view that as the novelty.
Response: I am thankful to the reviewer for this valuable suggestion. I do agree that by emphasizing on the clinical relevance, the contribution and quality of the paper will enhance. I have also revised the title as required to increase the impact of the paper.

Comment 2: I would move Table 1 into the text.
Response: I thank for this suggestion. I have repositioned Table 1 and now it appears immediately after its first reference in the text. This will improve the readability and also the structural flow.

Comment 3: This study needs to be cited: “A comparison study clearly demonstrates the superiority of pre-trained domain-specific AI models.”
Response: Thank you very much for pointing this out. As per suggestion, I have revised the statement and also clarified well. It is now supported by existing literature and I have aligned it with proper citations which are already in the paper (e.g., Raghu et al., 2019; Litjens et al., 2017).

Comment 4: The manuscript would benefit from defining what pretraining is and going through more of the technical details before assuming familiarity on the subsequent modality discussions. I would recommend adding a section on ML and pre-training for medical imaging before going into the different modalities.
Response: I am in full agreement with this important suggestion. For this I have added a totally new section titled “Foundations of Pre-Training in Medical Imaging” in between Introduction and Methodology sections. The concept of pre-training has been introduced and it explains its role in deep learning workflows, providing necessary context for the modality-specific discussions that follow.

Comment 5: It would also benefit the article to specifically discuss why AI models fail without pre-training in each specific case and what pretraining can be done to improve them. Some technical details help.
Response: I am also thankful to the reviewer for this valuable suggestion. Accordingly, I have revised the modality-specific sections (Chest X-ray, CT, MRI, and Ultrasound), discussing the following in a more explicit manner
(i) the limitations of general-purpose foundational AI models in each modality.
(ii) how domain-specific pre-training addresses these limitations.
Additional technical explanations have been properly incorporated regarding feature learning, noise robustness, and modality-specific characteristics paper to strengthen the discussion.

Smaller Edits

Comment 1: In the abstract, remove “To be precise” to increase flow.
Response: This phrase has been removed from the abstract to improve clarity and readability.

Comment 2: Should be: “Healthcare is one of the most critical domains because even the slightest of the errors in any kind of diagnosis can highly impact the safety of the patient and treatment planning; therefore, reliability is one of the most important factors for any AI system in the healthcare sector.”
Response: I have revised the sentence in the abstract to reduce syntactical errors and improve clarity while maintaining formal academic tone.

Comment 3: Delete “and are one of the ultra-critical and compounded parts of that.” Doesn't make sense. Also, one thing does use the plural verb are.
Response: This phrase has been removed as suggested to improve sentence clarity.

Comment 4: Should be challenges: “General foundational AI models are still used in spite of such challenges.”
Response: This has been corrected properly.

Comment 5: The word equates does not make sense here: “With a proper combination of existing ideas and elements, this review paper equates general purpose foundational AI models on the basis of four major imaging modalities, namely- chest X-ray, MRI, CT and ultrasound.”
Response: I have replaced the term “equates” with “examines and compares” to accurately show the meaning.

Comment 6: In many places you use the word “the”, where it shouldn't be used in English. Please use a grammar check to fix this.
Response: I have carefully checked the entire paper and corrected the improper article usage throughout.

Comment 7: Don't need “The last section, that is ultrasound imaging…”
Response: This phrasing has been corrected.

Comment 8: Delete whereas in “Whereas in contrast”.
Response: The redundancy is corrected.

Comment 9: Nothing was tested in this paper, everything was discussed: “Through the broad spectrum of the imaging modalities tested in this review paper”
Response: This has been corrected to properly reflect the nature of the paper.

Response to Reviewer 2

I am sincerely grateful to the reviewer for such a detailed thoughtful and constructive feedback. I greatly appreciate for recognising the strength of the paper, its conceptual framing, clinical relevance and cross-modality structure. As per suggestion I have revised the paper very carefully, addressing all concerns focussing on improving the methodological clarity, evidentiary rigor, writing precision, and structural organization.

Major Revisions (Research Methods Focus)

Comment 1: Clarify review methodology (study selection, inclusion/exclusion criteria, databases, etc.).
Response: I am in full agreement with this very important suggestion. Addressing this I have added a dedicated section named "Methods" describing clearly the review methodology. Specifically, I have included Databases used (PubMed, IEEE Xplore, arXiv), Keywords used for literature search, Inclusion and exclusion criteria, focus on peer-reviewed and widely cited studies, Explanation that this is a narrative review (not a formal meta-analysis), Description of how performance metrics were interpreted across studies.

Comment 2: Improve quantitative comparisons and avoid vague descriptors.
Response: Thank you very much for the valuable suggestion. As per suggestion, the tables and associate discussions has been revised, the clarity and precision has also improved. Specifically, reported metrics such as AUROC and Dice scores has been included wherever possible, clarified when comparisons are dataset-dependent, vague descriptors like "high" or "moderate" has been removed wherever unsupported. I further ensured that claims of performance differences are aligned with reported values. In addition to these, I have clarified that cross-study comparisons are limited due to dataset and evaluation differences.

Comment 3: Inconsistency in Table 2 (similar AUROC but claim of superiority).
Response: I appreciate this important observation. To avoid overstating superiority, I have revised the discussion, emphasising on Performance differences depend on dataset and evaluation conditions, Domain-specific models often show improved generalization and robustness rather than universally higher AUROC, at the same time ensuring consistency between the table and the interpretation

Comment 4: Avoid overgeneralization (e.g., "incorrect almost every time").
Response: I fully agree with this concern. Accordingly, I have revised all such statements reflecting on a more nuanced and evidence-based perspective. Specifically: softening strong claims, language now reflects "limited generalization" instead of "failure", Conditions under which general models underperform are explicitly described. This has improved the scientific accuracy and tone of the paper.

Structural and Clarity Improvements

Comment 5: Reduce repetition across sections.
Response: Redundancy reduced throughout the paper.

Comment 6: Improve writing precision and academic tone.
Response: I thank the reviewer for this suggestion, to improve Grammar and sentence structure, removing redundant phrases, replacing informal expression with academic language and to improve clarity and conciseness, I have thoroughly edited the paper Thais revision. Significantly these enhances the overall readability and professionalism of the paper.

Minor Suggestions

Comment 7: Clarify dataset statistics.
Response: Thank you for this feedback. I have clarified the dataset statistics in the paper.

Comment 8: Ensure consistency in citation formatting.
Response: Thank you for this feedback. I have formatted the citations properly throughout the paper.

Comment 9: Add a subsection on limitations of domain-specific pre-training.
Response: I have added the subsection on the limitations of domain-specific pre-training.

Author 100154 Second Review

Thank you for your careful revisions of your manuscript and attention to our suggestions. After reviewing your corrections, I recommend that you revise with minor revisions.

1. I thank you for more carefully refocusing the title and Intro to emphasize your exact application instead of a generic discussion of pretraining.
2. The most important edit I would like to re-emphasize from my original comments is that you should thoroughly describe the machine learning architectures used in your Foundations of Pre-Training section. This should emphasize a combination of images and equations, not text, since architectures can't be explained precisely in words. Please show how the training and pre-training processes work for, at least, a few example cases you describe in the literature.
3. I would still encourage you to more carefully state what makes the architectures work and not work in different situations. They don't always work. Why? When they do work, why? That leads to a critical understanding of the field.
4. The limitations of the domain-specific pretraining section should either be longer (there is much to say here) or be integrated into the Discussions section; it seems orphaned right now.
5. You should discuss more of the challenges and opportunities for these methods looking forward. What are competing methods (like metalworking and transfer learning)? How do they perform? How might these competing approaches perform in the future?

Final Decision: Accept with Minor Revisions

Thank you for the corrections you have submitted. I really appreciate the clarity and structure of your manuscript. The way you present and organise your discussion works well, and it reflects a strong grasp of the subject area. Your synthesis of the literature is clear and accessible, making the paper engaging to read. I particularly like how you guide the reader through each section in a logical and coherent way.

That said, there are a few areas where some refinement would further strengthen the paper. The methodology section needs to explain your approach in more detail. While you identify the work as a narrative review, it would be beneficial to outline your search strategy more explicitly, including the databases used, inclusion and exclusion criteria, and any timeframes applied. This would improve transparency and enhance the academic rigor of the paper.

You might also consider deepening the critical engagement with the literature. It might be also good to section your paper with a Literature Review chapter to focus your flow of thought. While the summaries are clear, it would add value to compare studies more directly, particularly in terms of datasets, evaluation approaches, and limitations. It might be worth adding a more consistent way of presenting performance metrics such as AUROC and Dice scores to allow clearer cross-study comparison.

In terms of writing, I really like the clarity in several sections; however, there are areas where the phrasing becomes repetitive or slightly informal. A focused language edit would help improve conciseness and maintain a consistent academic tone throughout. It might be also good to go through your language and de-AI is so that the flow of the content is clear and deep.

It might also be worth addressing the formatting issue in the ultrasound section, where part of the text is obscured by an overlapping image, as this affects readability.

Finally, you might consider expanding the conclusion to include more explicit and actionable future research directions, particularly around real-world implementation and data-related challenges.

I have reviewed the student's responses to the peer reviewers and the revised submission. I agree with both reviewers in that the paper is at a stage of **Accept with Minor revisions.**

Between the peer reviewers, they have picked up on most of what arose for me as well. My comments are primarily around the structure of the paper:

Abstract

- The abstract could be strengthened if it was structured ie using subheadings Background, Methods, Results, Conclusion

Introduction

- The aim of the study should be included at the end of the Introduction before going into Methodology. This would help the reader understand the purpose of the paper.
- It is a bit odd to have just the subheading "Foundations of Pre-Training in Medical Imaging" as the sole subheading in the Introduction. It seems unnecessary and should be removed, with these paragraphs just forming part of the Introduction.

Methodology

- What is missing is a more detailed methods of how the review was conducted with an appropriate search strategy ie what types of articles were included/excluded, time period of publication, English-language only? etc

Results

- Include a heading for Results so that we know when it begins.
- The use of tables in the paper is good as it summarises information well, however they are not referred to in-text, so it is unclear when the reader should be referring to them.
- Try to not use abbreviations in tables, and if you do, you need to define what they are.

Discussion

- It is good that a limitations section is included, however the purpose of these is usually to discuss the limitations of the methodological process eg generalisability of findings, small sample size etc, rather than the limitations of the topic you are discussing. It is advised that you should remove the heading "Limitations of Domain-Specific Pre-Training".

Overall

- Slight grammatical issues throughout including capitalisation of words that do not require it eg Medical Imaging, Chest X-ray, Computed Tomography, Magnetic Resonance Imaging etc.

Enhancing the clinical reliability of AI models in medical imaging through the use of domain-specific pre-training across modalities

Background: Artificial Intelligence (AI) has been integrated into many sectors, which has led to improvements in accuracy and efficiency. AI has notably improved performances in various tasks such as image classification, object detection, and segmentation, most of the time outperforming manual approaches in different controlled environments. Healthcare is one of the most critical domains because even the slightest of errors in any kind of diagnosis can highly impact the safety of the patient and treatment planning therefore, reliability is one of the most important factors for any AI system in the healthcare sector. Despite this, many AI models, which are primarily trained on natural images, are used in various healthcare applications, which limits their clinical reliability and generalizability**.**

Methods: This manuscript presents a narrative review of existing literature, comparing general-purpose foundation AI models and domain-specific pre-trained AI models across four major imaging modalities, namely, chest X-rays (CXR), magnetic resonance imaging (MRI), computed tomography (CT), and ultrasound. Relevant studies were identified through searches across major databases including PubMed, IEEE Xplore and arXiv, using predefined inclusion and exclusion criteria. This review is qualitative in nature and does not perform a formal meta-analysis.

Results: The AI models that are trained on domain-specific data consistently show elevated performances in clinically applicable tasks, as reported across multiple studies in terms of performance, robustness, and clinical applicability when compared to AI models pre-trained on natural image datasets.

Conclusion: Domain-specific pre-training will not only benefit, but will also be an indispensable element for developing reliable and clinically useful AI systems in the field of medical imaging. Future research should focus on multi-modal data integration, cross-institutional validation, real-world clinical testing, and comparison with alternative learning paradigms such as transfer learning and meta-learning.

Keywords: AI, medical imaging, domain-specific pre-training, chest X-ray, MRI, CT, COVID-19, ultrasound, human-AI collaboration

Introduction

Medical imaging is vital for modern clinical diagnosis, allowing non-disruptive imaging of anatomical structures and any abnormal bodily function of cells/tissues, required for treatment, prognosis, and disease mapping. Imaging modalities such as chest X-ray, computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound are widely used in the healthcare industry and are a critical component of the diagnostic workflow.

In recent years, Artificial intelligence (AI), particularly deep learning (DL), has been a significant tool for medical image analysis, allowing noticeable improvements in diagnostic accuracy, efficiency, and clinical progress (Litjens et al., 2017). The success of deep learning in general computer vision has quickened the adoption of general-purpose foundational AI models in medical imaging, most of which are pre-trained on large-scale natural image datasets.

These AI models have shown exceptional accomplishment across a wide range of general visual tasks, which motivated their shift to medical imaging applications. As medical images are modality-specific, they greatly differ from natural images and thus require trained and expert clinical data interpretation. Consequently, AI models trained on the basis of non-medical data often fail to show effective results in the clinical field or practice. This raises concern regarding safety, accuracy, and dependability (Kelly et al., 2019; Roberts et al., 2021).

Every imaging modality faces challenges and difficulties and thus demands domain-aware learning. In the case of chest X-rays, there are two-dimensional representations with overlapping anatomical structures that require refined layered interpretation of distinct regions of the lungs, cardiac profile, and delicate radiographic patterns for the identification of diseases such as pneumonia and tuberculosis. Magnetic resonance imaging (MRI) offers a superior soft-tissue contrast through multiple phase sequences (e.g., T1, T2, T1ce), thereby enabling a comprehensive understanding of neurological and musculature abnormalities, necessitating an understanding of complicated tissue-specific magnitude variations.

General foundational AI models are still used in spite of such challenges. With the help of large-scale surveys, it has been found that general foundational AI models are often not accurate and have inadequate key features like clinical bases and lead to wrong explanations of anatomical features (Tajbakhsh et al., 2016; Litjens et al., 2017). This review emphasizes the need for domain specific pre-trained fundamentals for more reliable medical imaging AI models. With a proper combination of existing ideas and elements, this review paper studies and analyzes general purpose foundational AI models on the basis of four major imaging modalities, namely: chest X-ray, MRI, CT and ultrasound.

A comparison across existing literature demonstrates the advantages of domain-specific pre-training in improving model robustness and clinical relevance (Litjens et al., 2017; Raghu et al., 2019).

This study aims to evaluate the role of domain-specific pre-training in the improvement of the clinical reliability of AI models across major imaging modalities and also aims to identify the key limitations and directions for future research in this field.

The process of training an AI model on a large dataset to learn general feature representations before fine-tuning it on a specific task is referred to as “Pre-Training”. In medical imaging, general-purpose pre-training on foundational AI models (e.g. ImageNet) mostly provides an insubstantial benefit due to the fundamental differences between natural and medical images, which include modality specific intensity distributions and anatomical structures (Raghu et al., 2019).

Domain-specific pre-training in AI models addresses this limitation by training the models directly on suitable medical datasets, thereby enhancing the learning of clinically applicable features and improving robustness to modality-specific noise and variability. This study builds on the framework to evaluate how pre-training strategies guide model reliability across various imaging modalities.

This paper aims to systematically evaluate that how domain-specific pre-training influences the clinical reliability of AI models across major imaging modalities, while also identifying key limitations, failure cases, and future research directions which are necessary for real-world deployment.

Model architectures and the pre-training mechanisms in medical imaging

In the context of a narrative review, mathematical or architectural derivation in detail is beyond the scope. The crucial thing is to understand conceptually the interaction of the different model architectures with pre-training strategies. Most of the AI models that are used in medical imaging are based on convolutional neural networks (CNNs) and also their extensions which include 3DCNNs and encoder-decoder architectures such as U-Net. As far as capturing spatial hierarchies, contextual information, and modality-specific patterns, these architectures differs.

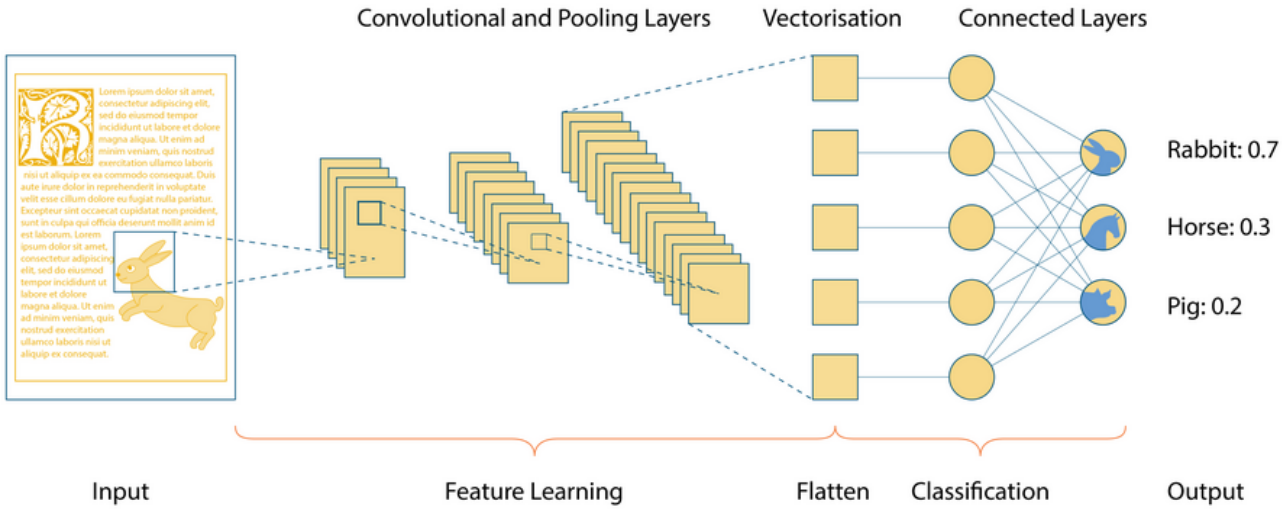


Figure 1

(Figure 1 Standard convolutional neural network (CNN) architecture illustrating hierarchical feature extraction and classification *(*adapted from Wikimedia Commons, CC BY-SA 4.0)).

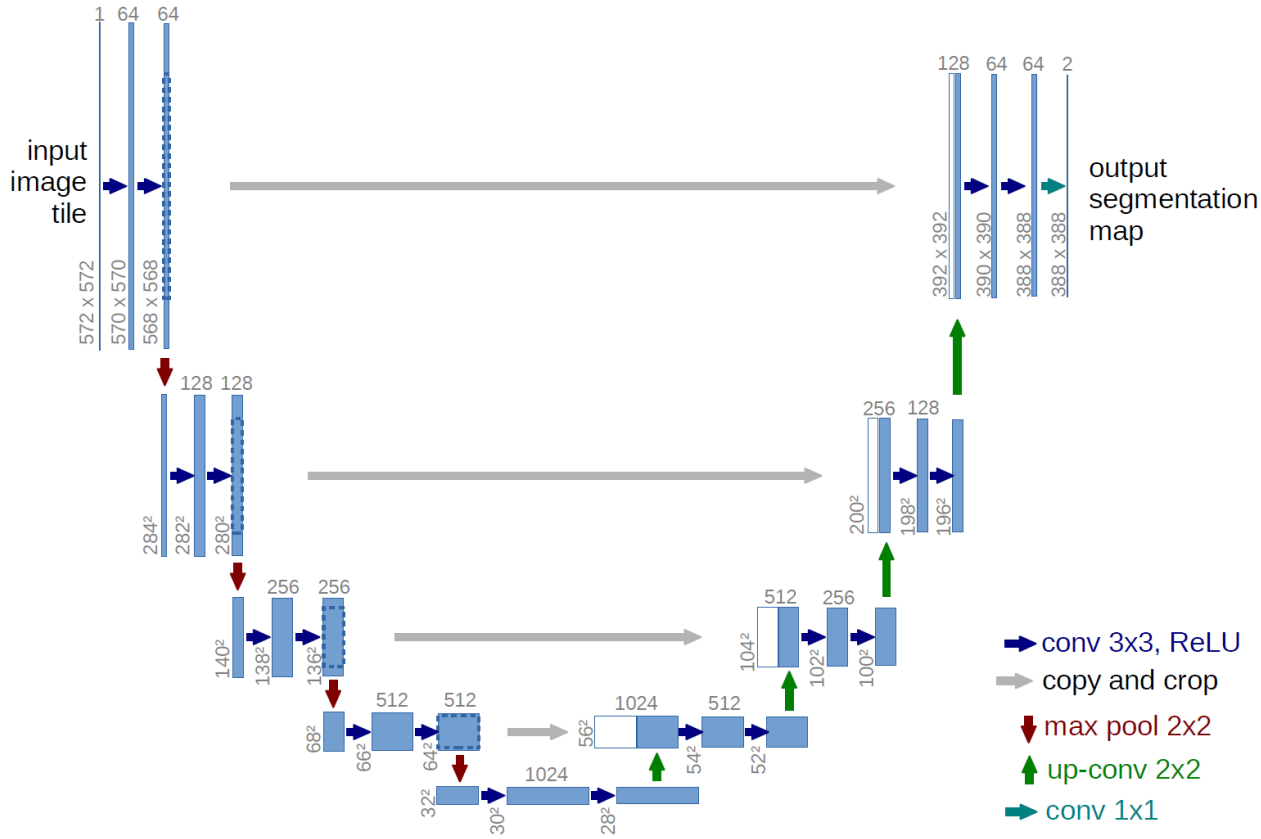


Figure 2

(Figure 2 shows the U-Net architecture, an encoder–decoder network designed for biomedical image segmentation, where the contracting path captures context and the expanding path enables precise localization through skip connections (Ronneberger et al., 2015)).

The general-purpose foundational AI models which are pre-trained on natural image datasets like ImageNet, learn only the low-level features like edges, texture and the basic shapes. Although these features are able to transfer to some extent, they are not sufficient for capturing clinically relevant patterns like tissue density variations, anatomical structures, and pathology-specific signals. This limitation is mainly evident in modalities like MRI and CT, in which the intensity distribution and spatial relationships are fundamentally different from natural images.

On the other hand, the domain-specific pre-training enables models to learn the feature representations which are directly aligned with medical imaging features. For example, 3D CNNs that are trained on volumetric CT data are able to capture inter-slice special continuity, while MRI-specific models can learn multi-sequence dependencies across T1, T2 and contrast-enhanced images. Similarly, the ultrasound-specific models become robust to speckle noise and operator variability when exposed to domain-relevant data distributions.

The effectiveness of these architectures depends strongly upon the alignment between the training data and the target clinical task. The case where the pre-training data closely matches the modality and task distribution, the models tend to perform well. This allows them to learn meaningful representations. Conversely, when there is a mismatch in the domain, the performance degrades which leads to poor generalisation, increased false positives and unreliable clinical predictions. Therefore, it is the combination of architecture design and domain-specific pre-training which determine the success of AI systems in medical imaging and not the architecture alone.

Methodology

This study is a narrative review that synthesizes existing literature on the role of domain-specific pre-training of AI models in medical imaging. Relevant studies were identified through the different searches conducted on databases such as PubMed, IEEE Xplore, and arXiv. Keywords used were, “Medical Imaging”, “Transfer Learning”, “Domain-Specific Pre-Training”, “MRI Deep Learning”, “CT Segmentation”, “Ultrasound AI”, and “Chest X-ray Deep Learning”. Inclusion criteria included the peer-reviewed publications and some widely cited research papers that provided empirical evaluation of AI models in various medical imaging tasks. Studies which compared domain-specific pre-training with general-purpose pre-

training were prioritized. The selection process prioritized peer-reviewed publications and widely cited research papers that provided empirical evaluation of AI models in medical imaging tasks such as image classification, detection, and object segmentation. Studies which compared domain-specific pre-training with general-purpose pre-training were included and discussed on the basis of modality-specific challenges in depth in this paper. Papers which lacked methodological clarity or empirical validation were excluded. Exclusion criteria included the studies which lacked empirical validation or had unclear methodologies. Studies which were published between 2015 and 2023 were prioritized to reflect recent developments in deep learning. This review is qualitative in nature and does not perform a formal meta-analysis; instead, performance metrics such as AUROC, Dice scores, sensitivity, and specificity are interpreted as mentioned in the original studies, with proper consideration of the dataset differences and experimental conditions.

Results

The findings across the reviewed studies indicated that domain-specific pre-training consistently improved performance, robustness, and clinical applicability of the AI models across the various imaging modalities. Table 1 summarizes the key imaging modalities, associated AI tasks, and challenges.

Table 1: Overview of various Imaging Modalities

Modality	Key AI Tasks	Example Domain Specific Models	Challenges Addressed
Chest X-ray (CXR)	Pneumonia detection	CheXNet	Low resolution and overlapping structures
Computed Tomography (CT)	Lesion detection, cancer segmentation	3D U-Net, multi-view CNN	False positives, scanner variability
Magnetic Resonance Imaging (MRI)	Tumor segmentation, classification	nnU-Net, MRI-pretrained CNN	High soft tissue contrast and multi-sequence data
Ultrasound	Organ segmentation, echocardiography analysis	US-specific CNNs	Speckle noises and operator dependencies

(Performance comparisons across studies should be interpreted cautiously, as variations in datasets, evaluation protocols, and class distributions significantly influence reported AUROC and Dice scores; Abbreviations: CNN - Convolutional Neural Network; AUROC - Area Under the Receiver Operating Characteristic Curve)

Hardware workings of the various medical imaging modalities and their limitations

The chest x-ray imaging is primarily based on the photons passing through the tissues of various densities present. While this method is fast and cost-effective, there are some limitations to it as well, which is that the X-rays have many 2D projections with multiple overlapping anatomical structures, which limit the depth perception and also the sensitivity factors to subtle pathologies such as early COVID-19 infiltrates and other kinds of severe lung diseases.

Computed tomography or CT, is widely used currently in the healthcare industry. The CT scans reconstruct multiple X-ray projections into volumetric 3-dimensional representations using various advanced tomographic reconstruction algorithms and therefore it offers superior resolution and also shows a proper and a better density variation, but it also has a major limitation that is it exposes the patient to a higher amount of radiation doses which is extremely harmful.

Magnetic resonance imaging or MRI relies completely on the nuclear magnetic resonance principles to capture the tissue-specific contrast through variations in the relaxation timings (T1, T2) and the pulse sequences. MRI also provides an exceptionally soft tissue contrast but on the other hand it suffers from extremely high cost, long scan times and also scanner-dependent variability.

Ultrasound imaging, uses high frequency sound waves to generate real-time images. The advantage of this is that it is safe, portable and cheaper compared to the other imaging modalities but it is highly sensitive to speckle noises and operator dependency.

Chest X-Ray Imaging

One of the most widely used diagnostic modalities commonly used in clinical practices is the chest X-ray (CXR). This is generally used for the detection and also monitoring of pulmonary and cardiovascular diseases. Since the cost is low, the outcome is prompt, and it is also broadly available, chest radiology is considered to be the primary and the first imaging tool for detecting conditions such as pneumonia, tuberculosis, lung cancer, COVID-19, pleural effusion, heart failure and other such things.

However, chest X-rays are not free from complex analytical challenges. Plenty of such challenges are present for automated systems and AI models. One of the important reasons for this is that since the images are two-dimensional (2D), the overlapping anatomical structures and subtle intensity variations require expert interpretation. X-ray analysis is therefore very complex, and various studies show that the AI models pre-trained on domain-specific data mostly surpass general-purpose foundational AI models in clinically relevant tasks when evaluated under comparable conditions (Raghu et al., 2019; Litjens et al., 2017).

This type of domain-specific pre-trained model becomes familiar with clinical terminologies, even the critical, advanced and complex ones. Not only this, but these models also become familiar with meaningful radiographic patterns, for example, lung-field asymmetries, interstitial markings and subtle opacities. These were generally missed or misinterpreted in the case of AI models that are pre-trained on natural images. General AI models that are pre-trained on natural images like ImageNet have difficulty perceiving subtle disease-specific patterns, due to negligible visibility to relevant medical imaging features (Zech et al., 2018).

One of the most prominent examples of such training in the case of chest X-ray imaging is *CheXNet,* proposed by *Rajpurkar et. al. (2017)*. Beyond pneumonia, various recent studies and literature have demonstrated the effectiveness of the usage of domain-specific AI models in detecting COVID-19 related things, tuberculosis cavitation and lung nodules as well, which are often subtle and easily missed by the general-purpose AI models. Since it was trained on large-scale expertly labelled X-ray datasets, CheXNet achieved state-of-the-art performance on multiple thoracic diseases, including pneumonia, with AUROC scores typically in the low-to-mid 0.80s on the ChestX-ray14 test set, comparable to radiologist-level performance (Rajpurkar et al., 2017).

However, AUROC alone is sometimes not sufficient for cross study comparison as the datasets, disease prevalence, and evaluation protocols differ greatly. It was also found that CheXNet illustrated improved diagnostic performances in several reviewed parameters as compared to the general convolutional neural networks (CNNs). On the other hand, the performances of general-purpose CNNs pre-trained on natural images were stated to show limited generalizability in certain clinical settings (Zech et al., 2018). Zech et al. (2018) also showed that the AI models that were trained on chest X-rays from a single institution were unable to generalize to external datasets. Especially when pretrained on non-medical images.

From these findings, we can conclude that domain-specific pre-training will not only give more accurate results but will also reduce data bias to a large extent, therefore increasing robustness and reliability in real-world clinical environments. There are more recent studies that comparatively evaluated and found that the models pretrained on large-scale chest X-ray datasets consistently performed much better than general-purpose vision models of comparative size in both the classification and segmentation tasks (Irvin et al., 2019). The AI models that are pretrained on chest radiography are much aligned with radiological reasoning. Due to this the findings are better, and false positives are much less. This alignment is very important; since these are related to health, a small misclassification may lead to serious clinical consequences.

These findings indicate that extensive evidence has been confirmed in both foundational and recent studies that domain-specific pre-training is a significant factor on which AI performance in the case of chest X-ray imaging depends. These findings very clearly state that modality-aware pre-training is very important and essential for AI systems in chest radiography to be clinically trustworthy, and this provides a strong motivation to extend these domain-specific training strategies across other medical imaging modalities. A large chest radiography dataset contains 224,316 images from 65,240 patients (Irvin et al., 2019). The deep learning models which were trained on chest X-ray images specifically achieved a higher multi-pathology classification performance compared to the general Foundation AI models.

Table 2: Chest X-ray AI Model Comparison

Model	Pre-training	Task	Performance	Notes
CheXNet	Chest X-ray dataset	Pneumonia classification	AUROC ≈ 0.82–0.85	Evaluated on ChestX-ray14 dataset
CNN	ImageNet	Pneumonia classification	Performance varies across datasets; often lower generalization in clinical settings	Generalization issues reported (Zech et al., 2018)
CheXNet	Chest X-ray dataset	Multi-disease detection	Performance depends on dataset and labeling quality	Supported by multiple studies

(As shown in Table 2, domain-specific models such as CheXNet demonstrate improved robustness and generalizability compared to general-purpose CNNs, particularly under cross-institutional evaluation settings; Performance comparisons across studies should be interpreted cautiously, as variations in datasets, evaluation protocols, and class distributions significantly influence reported AUROC and Dice scores; Abbreviations: CNN - Convolutional Neural Network; AUROC - Area Under the Receiver Operating Characteristic Curve).

Computed Tomography (CT) Imaging

Computed tomography imaging plays a vital role in day-to-day clinical workflows, especially in oncology and trauma assessment tests. CT scan imaging is also widely used in the detection of pulmonary embolism, stroke tests, intracranial hemorrhage identification and COVID-19 tests as well. CT scans provide high-resolution three-dimensional (3D) volumetric data which capture very finely grained anatomical structures and details across a wide range of slices. These characteristics demand AI models which are capable of properly understanding the complex 3D spatial relationships, tissue density variations, and subtle pathological patterns.

Computed tomography (CT) imaging represents one of the most complex modalities for reliable automated analysis. General-purpose foundational AI vision models which are trained on two-dimensional natural images, most of the time struggle to interpret the volumetric CT data effectively. These models lack intrinsic mechanisms which are necessary for learning the spatial continuity across various slices and may sometimes fail to capture clinically relevant 3D contextual data.

In contrast, the AI models trained on volumetric CT data demonstrate highly improved sensitivity and robustness in the clinical detection and segmentation tasks (Çiçek et al., 2016; Setio et al., 2017). Previous studies demonstrate that CT-specific models which are trained on volumetric data improve cancer detection sensitivity and lesion localization to a great extent when compared to 2D vision models pretrained on natural images (Çiçek et al., 2016; Setio et al., 2017). Several studies show high improvements in sensitivity for lesion detection when CT-specific 3D architectures are used in comparison to 2D vision models (Setio et al., 2017).

CT-specific domain-trained models have been demonstrated to reduce false positives in lesion detection tasks, which is an extremely important and crucial factor in minimizing unnecessary follow-up procedures (Setio et al., 2017). These findings show the disadvantages of the general-purpose foundational AI models in handling volumetric medical data and also highlight the necessity of modality-aware pre-training.

Another major advantage of CT-specific domain pre-training is in its ability to account for scanner-related variability and reconstruction artefacts. The CT images vary a lot, depending a lot on the scanner manufacturers, the radiation dose protocols and the reconstruction algorithms. The pre-training of AI models with specific domain data enables the AI models to learn invariant features properly and in depth so that they can generalize across such variations, whereas general-purpose vision models often wrongly interpret reconstruction artefacts as pathological features and hence limit their clinical reliability (Litjens et al., 2017; Roberts et al., 2021).

Overall, this research and these studies demonstrate that CT imaging shows fundamental shortcomings in general-purpose

foundational vision AI models, and it also strongly supports the adoption of domain-specific data pre-training methods. For complex and critical applications like cancer detection and trauma diagnosis, the domain-aware CT AI models are extremely needed for achieving a clinically reliable and trustworthy AI performance.

Table 3: CT AI Performance

Model	Pre-training Type	Task	Performance	Notes
3D CNN	CT-specific dataset	Lesion detection	Higher sensitivity reported in controlled studies	Captures volumetric context
2D CNN	ImageNet	Lesion detection	Lower sensitivity in volumetric tasks	Limited 3D understanding

(As summarized in Table 3, CT-specific models such as 3D CNNs outperform 2D models due to their ability to capture volumetric spatial relationships; Performance comparisons across studies should be interpreted cautiously, as variations in datasets, evaluation protocols, and class distributions significantly influence reported AUROC and Dice scores; Abbreviations: CNN - Convolutional Neural Network; AUROC - Area Under the Receiver Operating Characteristic Curve).

Magnetic Resonance Imaging (MRI)

Magnetic Resonance Imaging has emerged as one of the most important imaging tools in modern-day medical diagnosis, especially in the fields of neurological imaging, oncology, and musculoskeletal imaging. Unlike other medical imaging tools such as X-ray technology-based medical imaging modalities, the MRI image has very high-resolution volumetric data with very subtle contrast in soft tissues that are demonstrated through appropriate image acquisition protocols such as the T1-weighted image, the T2-weighted image, and the contrast-acquired image (T1ce image). This very multi-sequence and high-dimensional nature of the MRI data makes it in particular less acceptable for direct transfer learning from the natural image datasets.

The analysis of an MRI image becomes more complex for a general AI system, this is because there is a large variation in the intensities specified to MRI image capturing tools which is again combined with a large amount of anatomical details which were absent in natural images. Due to this, MRI serves as a critical platform for evaluation of the effectiveness of domain-specific pre-training in medical imaging models. The transfer learning from the natural image datasets most of the times provide a slight benefit in capturing the modality-specific intensity variations which are inherent to MRI data (Raghu et al., 2019).

A growing body of the existing literature continuously shows that the pre-trained AI models, those that are pre-trained on specific MRI datasets are much better than the general-purpose foundational AI models of same sizes across a wide range of tasks including tumor detection, classification and segmentation. Raghu et al. (2019) elaborately investigated the effectiveness of features learnt from the datasets which contained natural images in medical imaging tasks and demonstrated that the ImageNet pre-training provided very limited benefits for applications which were based on magnetic resonance imaging.

Their findings showed that when the AI models of the similar sizes were compared, the AI models which were pretrained directly on the MRI datasets consistently outperformed the ImageNet pretrained counterparts across medical tasks such as tumor detection and tumor segmentation. This limitation comes out most of the time because the MRI data contains a lot of modality-specific complex features and details, such as tissue contrast mechanisms and anatomical structures, which are definitely not present in the natural image datasets. Therefore, the general-purpose foundational AI models find it very difficult and complex to learn these complicated and meaningful MRI details and representations without proper domain exposures.

Studies support these findings by demonstrating that the AI models trained directly on MRI data achieve a greater and a better tumor segmentation performance across multiple sequences which include T1, T2, and also the contrast-enhanced MRI when compared to the AI models which are pretrained on natural images (Isensee et al., 2021; Bakas et al., 2018).

This advantage in the performance has been continuously observed on the public brain tumor segmentation benchmarks such as BRATS.

Across the various public brain tumor segmentation benchmarks, MRI-specific training strategies, such as nnU-Net, have been reported to achieve Dice similarity coefficients around or above 0.80 on the BRATS benchmarks, depending on the task and validation split (Isensee et al., 2021), while the AI models relying mainly on ImageNet pre-training consistently show a lower segmentation accuracy (Raghu et al., 2019; Isensee et al., 2021). This throws light on the limitations of the ImageNet based pre-training for the MRI-specific segmentation tasks.

These differences in the performances are particularly significant in clinical matters, where proper and precise tumor boundary demarcation directly influences the treatment planning and the diagnostic outcomes. In addition to these segmentation tasks, the MRI-specific foundational AI models have also demonstrated a much better performance in classification tasks like tumour grading and the detection of diseases. These tasks include the stroke outcome prediction as well.

Multiple studies and literature have reported that the MRI-specific representational learning improves the disease classification and also the tumor grading performances when compared to the general-purpose foundational vision AI models, especially when the model sizes are the same (Raghu et al., 2019; Litjens et al., 2017). These improvements show the importance of AI models learning MRI-specific anatomical priors, which the general vision AI models often fail to capture effectively and properly. Other important and crucial factors influencing the MRI performances are scanner variability and acquisition heterogeneity. MRI images vary greatly across the scanner manufacturers, field strengths and patient demographics. Domain-specific pre-training with proper and wide MRI datasets exposes the AI models to the scanner and acquisition variability during learning, which leads to improved robustness and cross-institutional generalization, in contrast, the general-purpose models often fail to adapt reliably and effectively to the unseen MRI distributions (Raghu et al., 2019; Roberts et al., 2021). Such robustness is extremely crucial for the safety and the reliability factor of clinical deployment of the MRI based AI models.

Overall, the extensive empirical evidence confirms that the MRI-based tasks substantially benefit from domain-specific pre-training with MRI-specific data when the AI models of the similar sizes are compared. These findings properly highlight that the MRI's modality-specific complexity requires proper and appropriate learning strategies and that the reliable clinical deployment of AI in MRI cannot rely solely on the general-purpose vision pre-training.

Table 4: MRI AI Performance

Model	Pre-training Type	Task	Dice Score / AUROC	Notes
nnU-Net	MRI dataset	Brain tumor segmentation	≈ 0.80+	BRATS benchmark (Isensee et al., 2021)
CNN	ImageNet	Brain tumor segmentation	Typically, lower than MRI-specific models	Limited transferability

(Table 4 highlights that MRI-specific models outperform ImageNet-pretrained models due to their ability to learn modality-specific tissue contrast and multi-sequence dependencies; Performance comparisons across studies should be interpreted cautiously, as variations in datasets, evaluation protocols, and class distributions significantly influence reported AUROC and Dice scores; Abbreviations: CNN - Convolutional Neural Network; AUROC - Area Under the Receiver Operating Characteristic Curve).

Ultrasound Imaging

Ultrasound imaging, due to its non-invasiveness, real-time acquisition and low cost, is being widely adopted across a wide range of clinical domains (e.g., abdominal imaging, obstetrics and cardiology). Other important applications include vascular assessment and echocardiography assessments as well. However, ultrasound imaging carries a number of complex and unique challenges, e.g., low signal-to-noise ratio, speckle noises, operator dependency, and considerable variability in image qualities. The general foundational vision AI models which are trained on natural image datasets may misinterpret speckle noises and acquisition artefacts as meaningful features, thereby leading to unreliable predictions (Xie et al., 2018;

Chen et al., 2019).

On the other hand, domain-specific pre-training on large-scale ultrasound data enables AI models to learn robust clinically relevant patterns despite the substantial noise and operator variability (Xie et al., 2018; Chen et al., 2019). Multiple surveys in the literature have shown that deep learning models pre-trained on domain-specific ultrasound data consistently yield strong performance compared to general foundational AI models with comparable parameter sizes (Xie et al., 2018; Chen et al., 2019).

The parameters such as accuracy, f1 score and robustness under the noise consistently favour the domain trained AI models. Specifically, in the noisy and operator-dependent imaging condition, AI models pre-trained on domain-specific pre-training data consistently outperform the general-purpose foundational AI models of comparable sizes (Xie et al., 2018; Chen et al., 2019).

These differences are very critical when AI is taken and applied under real-world clinical settings, especially when ultrasound images vary significantly across devices and patient populations. This significant performance gap between general-purpose foundational vision AI models and ultrasound-specific AI models accentuates the need for modality-aware training in AI models in the healthcare field for trustworthy, accurate, and reliable results.

Table 5: Ultrasound AI Performance

Model	Pre-training Type	Task	Performance	Notes
CNN	Ultrasound dataset	Clinical tasks	Improved robustness under noise conditions	Domain adaptation beneficial
CNN	ImageNet	Clinical tasks	Performance sensitive to noise and variability	Less robust

(As shown in Table 5, domain-specific ultrasound models demonstrate improved robustness to noise and operator variability compared to general-purpose models; Performance comparisons across studies should be interpreted cautiously, as variations in datasets, evaluation protocols, and class distributions significantly influence reported AUROC and Dice scores; Abbreviations: CNN - Convolutional Neural Network; AUROC - Area Under the Receiver Operating Characteristic Curve).

Discussion

Through the broad spectrum of the imaging modalities discussed in this review paper, one could notice a similar trend suggesting that AI models which are pre-trained on domain-specific medical data, often tend to be more powerful and reliable when contrasted with the foundational AI models. Various findings from medical imaging studies suggest that modality specific details and pathologies are most of the times absent in the natural image datasets due to which they may result in diminished performance in certain clinical tasks. Hence, we may conclude that the general-purpose foundational AI models rely on dataset-specific patterns and shortcuts to a great extent in generalizing the practical applications in the healthcare sector (Tajbakhsh et al., 2016; Raghu et al., 2019; Zech et al., 2018; Oakden-Rayner, 2020; Roberts et al., 2021). Sometimes, the general vision AI models may find it complex in terms of variations in image acquisition conditions, image reconstruction algorithms, and patient demographics as well. These variances are mostly better handled through domain-specific pre-trainings in AI models due to which more reliable interpretations are developed.

The domain-specific image datasets have also shown great future scope to improve. There are a few largely used image datasets in the domain-specific area which are not fully certified or verified in a correct manner by radiologists. Hence, residual variances may remain (Rajpurkar et al., 2017). However, a few image datasets lack a variety of demographic populations due to which AI-related systems dedicated to healthcare may raise concerns in terms of fairness and reliability (Oakden-Rayner, 2020; Roberts et al., 2021). To overcome these constraints, more interactions are needed between humans, doctors and AI researchers. Domain-specific AI models are not intended to replace doctors or radiologists. Instead, they serve as efficient tools to enhance human-AI collaboration. When these models are correctly developed with accurate training and used practically, they can act as a reliable second reader. This can help relieve radiologists from heavy workloads and aid in sound clinical decision-making. However, using radiologist-verified datasets is crucial for improving

fairness, safety, and reliability. The well-trained medical AI models can support radiologists by significantly reducing their workload and also acting as reliable readers. To understand the environments under which these AI models give clinically safe and trustworthy diagnostics results is extremely necessary for the application of AI in real world clinical practices.

Though there are substantial advantages of domain-specific pre-training, there are certain limitations as well. One major challenge is the limited availability of large-scale, high-quality annotated medical datasets, because expert annotation is both time-consuming and also very expensive. Moreover, many existing datasets lacks sufficient demographic diversity, which may lead to bias and limit the generalizability of AI models across different populations (Roberts et al., 2021; Oakden-Rayner, 2020). Furthermore, variability in imaging protocols, scanner types, and acquisition settings across institutions can affect model robustness despite domain-specific pre-training. The computational costs related to training large-scale domain-specific models also remain high, thereby posing practical constraints for widespread adoption.

Comparing with alternative approaches like transfer learning and emerging techniques like meta-learning, the domain-specific pre-training offers improved feature alignment, but may lack flexibility when they are applied to unseen tasks or modalities. Transfer learning enables faster adaptation with limited data, while meta-learning focuses on rapid generalization across tasks. However, these approaches may not capture modality-specific characteristics as efficiently as the domain specific pre-training. Future research should therefore focus on and explore hybrid strategies that combine domain-specific learning with adaptable frameworks, this will improve scalability and generalization.

Conclusion

This review clearly demonstrates that the need for domain-specific pre-training is a significant necessity for the reliable AI system for medical images. Future research should focus on multi-modal data integration, cross-institutional validation, improved dataset diversity, and real-world clinical deployment of AI systems. The trained AI systems for modality-aware datasets have demonstrated a reliable and phenotypically accurate outputs than the foundational vision AI systems for general applications. The outcomes also represent the significance of dependability for medical applications. The dependence on general vision AI systems provides a clearer understanding of the constraints of general vision AI systems for healthcare applications. The application of domain-specific Pre-trained AI systems would provide improved human-AI collaboration to align human behavior with better outcomes. Further research work would be required in multi-modal integration for successful utilization of AI systems for healthcare applications. Finally, the development of precise and reliable AI systems would be required for healthcare applications to incorporate understanding of the domain. Future work should also focus on systematic comparisons between domain-specific pre-training, transfer learning, and meta-learning frameworks so as to better understand their relative strengths and their limitations in the real-world clinical applications.

Acknowledgement

The author of this paper acknowledges the contributions of all the researchers whose work formed the foundation of this review paper and also extends his sincere gratitude to his mentors and peers for their valuable guidance and feedback all throughout this research process.

Bibliography

1. 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.
2. Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems, 32.
3. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225.
4. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine, 15(11), e1002683.
5. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019, July). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 590-597).
6. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods, 18(2), 203-211.

7. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., ... & Jambawalikar, S. R. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629.
8. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning?. IEEE transactions on medical imaging, 35(5), 1299-1312.
9. Setio, A. A. A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., Van Riel, S. J., ... & Van Ginneken, B. (2016). Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE transactions on medical imaging, 35(5), 1160-1169.
10. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016, October). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In International conference on medical image computing and computer-assisted intervention (pp. 424-432). Cham: Springer International Publishing.
11. Roth, H. R., Lu, L., Liu, J., Yao, J., Seff, A., Cherry, K., ... & Summers, R. M. (2015). Improving computer-aided detection using convolutional neural networks and random view aggregation. IEEE transactions on medical imaging, 35(5), 1170-1181.
12. Smistad, E., Falch, T. L., Bozorgi, M., Elster, A. C., & Lindseth, F. (2015). Medical image segmentation on GPUs–A comprehensive review. Medical image analysis, 20(1), 1-18.
13. Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S. X., ... & Wang, T. (2019). Deep learning in medical ultrasound analysis: a review. Engineering, 5(2), 261-275.
14. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC medicine, 17(1), 195.
15. Oakden-Rayner, L. (2020). Exploring large-scale public medical image datasets. Academic radiology, 27(1), 106-112.
16. Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., ... & Schönlieb, C. B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence, 3(3), 199-217.
17. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Cham: Springer.

Enhancing the cClinical rReliability of AI models in mMedical iImaging through the use of dDomain-sSpecific pPre-Trainingtraining aAcross mModalities

Background: Artificial Intelligence (AI) has been integrated into many sectors, which has led to improvements in accuracy and efficiency. AI has notably improved performances in various tasks such as image classification, object detection, and segmentation, most of the time outperforming manual approaches in different controlled environments. Healthcare is one of the most critical domains because even the slightest of errors in any kind of diagnosis can highly impact the safety of the patient and treatment planning therefore, reliability is one of the most important factors for any AI system in the healthcare sector. Despite this, many AI models, which are primarily trained on natural images, are used in various healthcare applications, which limits their clinical reliability and generalizability.

Methods: This manuscript presents a narrative review of existing literature, comparing general-purpose foundation AI models and domain-specific pre-trained AI models across four major imaging modalities, namely, chest X-rays (CXR), magnetic resonance imaging (MRI), computed tomography (CT), and ultrasound. Relevant studies were identified through searches across major databases including PubMed, IEEE Xplore and arXiv, using predefined inclusion and exclusion criteria. This review is qualitative in nature and does not perform a formal meta-analysis.

Results: The AI models that are trained on domain-specific data consistently show elevated performances in clinically applicable tasks, as reported across multiple studies in terms of performance, robustness, and clinical applicability when compared to AI models pre-trained on natural image datasets.

Conclusion: Domain-specific pre-training will not only benefit, but will also be an indispensable element for developing reliable and clinically useful AI systems in the field of medical imaging. Future research should focus on multi-modal data integration, cross-institutional validation, real-world clinical testing, and comparison with alternative learning paradigms such as transfer learning and meta-learning.

Keywords: AI, medical imaging, domain-specific pre-training, chest X-ray, MRI, CT, COVID-19, ultrasound, human-AI collaboration~~Abstract¶~~

~~Artificial Intelligence (AI) has been integrated into many sectors, which has led to improvements in accuracy and efficiency. AI has notably improved performances in various tasks such as image classification, object detection, and segmentation, most of the time outperforming manual approaches in different controlled environments. Healthcare is one of the most critical domains because even the slightest of errors in any kind of diagnosis can highly impact the safety of the patient and treatment planning therefore, reliability is one of the most important factors for any AI system in the healthcare sector. Despite this, many AI models, which are primarily trained on natural images, are used in various healthcare applications, which limits their clinical reliability and generalizability. In this paper, the importance of domain-specific pre-training data across four major imaging modalities is thoroughly examined, namely, Chest X-rays (CXR), Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and Ultrasound. The AI models that are trained on domain-specific data consistently show elevated performances in clinically applicable tasks, as reported across multiple studies. The findings demonstrate that domain-specific pre-training will not only benefit but will also be an indispensable element for developing reliable and clinically useful AI systems in medical imaging and will also provide a clear direction for future research based on multi-modal data integration and real-world testing of AI in healthcare.¶~~

~~¶~~

~~**Keywords**: AI, Medical Imaging, Domain-Specific Pre-Training, Chest X-ray, MRI, CT, COVID-19, Ultrasound, Human-AI Collaboration¶~~

Introduction

Medical iImaging is vital for modern clinical diagnosis, allowing non-disruptive imaging of anatomical structures and any abnormal bodily function of cells/tissues, required for treatment, prognosis, and disease mapping. Imaging modalities

such as cChest X-ray (CXR), Ccomputed tTomography (CT), mMagnetic rResonance iImaging (MRI), and uUltrasound are widely used in the healthcare industry and are a critical component of the diagnostic workflow.

In recent years, Artificial iIntelligence (AI), particularly deep learning (DL), has been a significant tool for medical image analysis, allowing noticeable improvements in diagnostic accuracy, efficiency, and clinical progress (Litjens et al., 2017). The success of deep learning in general computer vision has quickened the adoption of general-purpose foundational AI models in medical imaging, most of which are pre-trained on large-scale natural image datasets.

These AI models have shown exceptional accomplishment across a wide range of general visual tasks, which motivated their shift to medical imaging applications. As medical images are modality-specific, they greatly differ from natural images and thus require trained and expert clinical data interpretation. Consequently, AI models trained on the basis of non-medical data often fail to show effective results in the clinical field or practice. This raises concern regarding safety, accuracy, and dependability (Kelly et al., 2019; Roberts et al., 2021).

Every imaging modality faces challenges and difficulties and thus demands domain-aware learning. In the case of chest X-rays, there are two-dimensional representations with overlapping anatomical structures that require refined layered interpretation of distinct regions of the lungs, cardiac profile, and delicate radiographic patterns for the identification of diseases such as pneumonia and tuberculosis. Magnetic rResonance iImaging (MRI) offers a superior soft-tissue contrast through multiple phase sequences (e.g., T1, T2, T1ce), thereby enabling a comprehensive understanding of neurological and musculature abnormalities, necessitating an understanding of complicated tissue-specific magnitude variations.

General foundational AI models are still used in spite of such challenges. With the help of large-scale surveys, it has been found that general foundational AI models are often not accurate and have inadequate key features like clinical bases and lead to wrong explanations of anatomical features (Tajbakhsh et al., 2016; Litjens et al., 2017). This review emphasizes the need for domain specific pre-trained fundamentals for more reliable medical imaging AI models. With a proper ¶
¶

combination of existing ideas and elements, this review paper studies and analyzes general purpose foundational AI models on the basis of four major imaging modalities, namely: cChest X-ray, MRI, CT and uUltrasound.

A comparison across existing literature demonstrates the advantages of domain-specific pre-training in improving model robustness and clinical relevance (Litjens et al., 2017; Raghu et al., 2019).

This study aims to evaluate the role of domain-specific pre-training in the improvement of the clinical reliability of AI models across major imaging modalities and also aims to identify the key limitations and directions for future research in this field.

The process of training an AI model on a large dataset to learn general feature representations before fine-tuning it on a specific task is referred to as “Pre-Training”. In medical imaging, general-purpose pre-training on foundational AI models (e.g. ImageNet) mostly provides an insubstantial benefit due to the fundamental differences between natural and medical images, which include modality specific intensity distributions and anatomical structures (Raghu et al., 2019).

Domain-specific pre-training in AI models addresses this limitation by training the models directly on suitable medical datasets, thereby enhancing the learning of clinically applicable features and improving robustness to modality-specific noise and variability. This study builds on the framework to evaluate how pre-training strategies guide model reliability across various imaging modalities.

This paper aims to systematically evaluate that how domain-specific pre-training influences the clinical reliability of AI models across major imaging modalities, while also identifying key limitations, failure cases, and future research directions which are necessary for real-world deployment.

Model architectures and the pre-training mechanisms in medical imaging

In the context of a narrative review, mathematical or architectural derivation in detail is beyond the scope. The crucial thing is to understand conceptually the interaction of the different model architectures with pre-training strategies. Most of the AI models that are used in medical imaging are based on convolutional neural networks (CNNs) and also their extensions which include 3DCNNs and encoder-decoder architectures such as U-Net. As far as capturing spatial hierarchies, contextual information, and modality-specific patterns, these architectures differs.

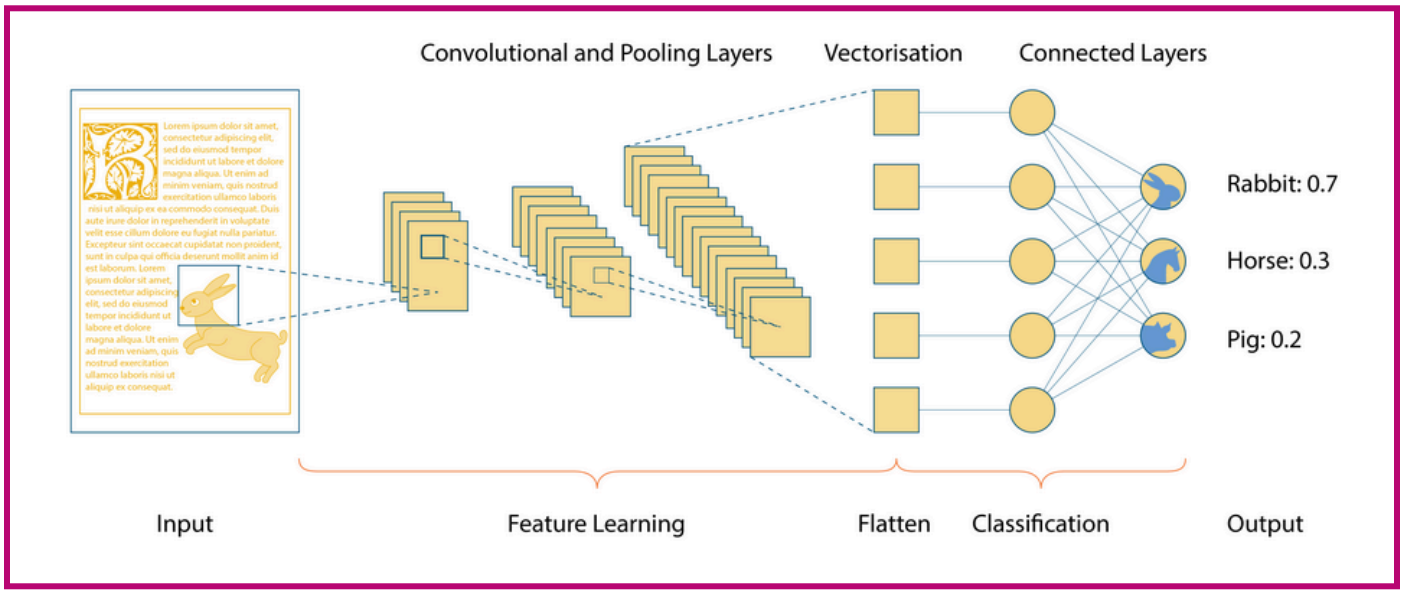


Figure 1

(Figure 1 Standard convolutional neural network (CNN) architecture illustrating hierarchical feature extraction and classification *(adapted from Wikimedia Commons, CC BY-SA 4.0)).*

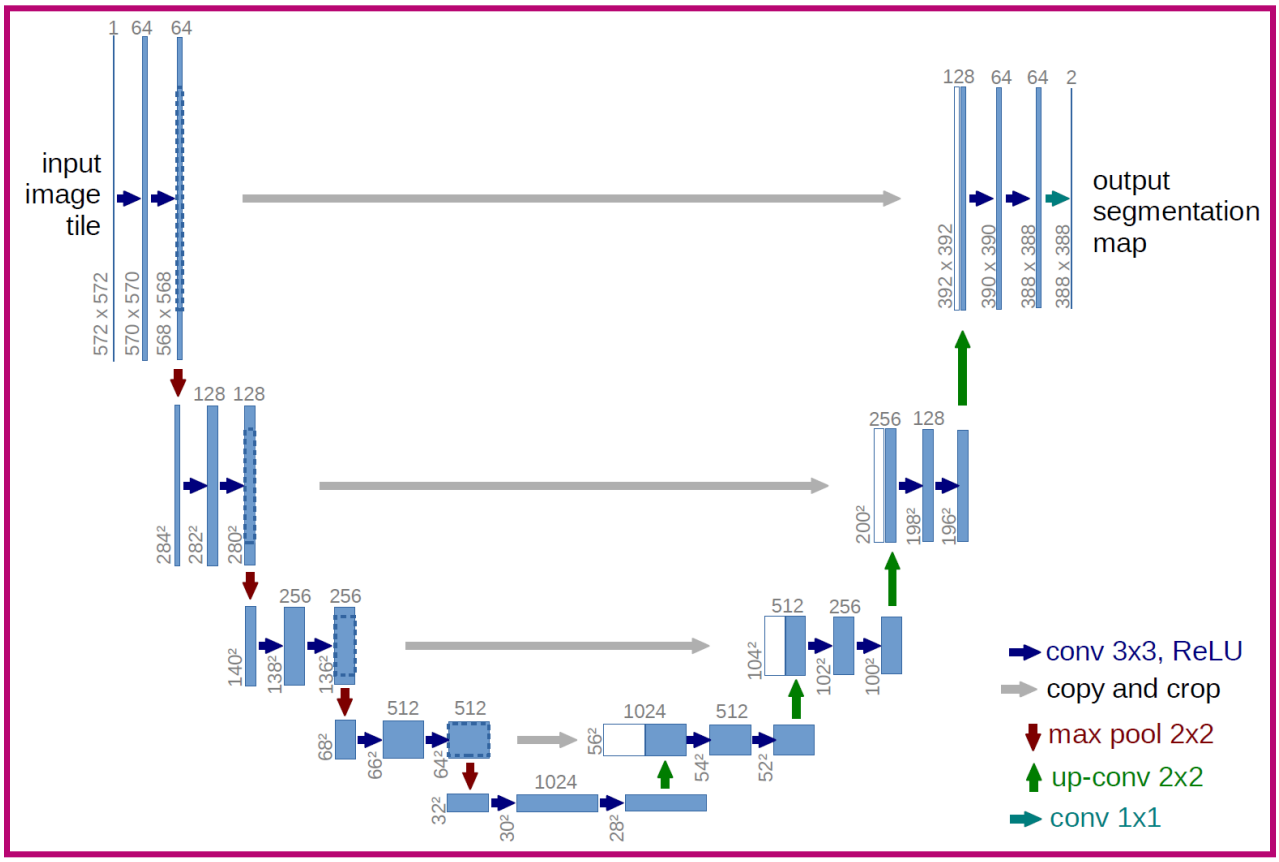


Figure 2

(Figure 2 shows the U-Net architecture, an encoder–decoder network designed for biomedical image segmentation, where the contracting path captures context and the expanding path enables precise localization through skip connections (Ronneberger et al., 2015)).
The general-purpose foundational AI models which are pre-trained on natural image datasets like ImageNet, learn only the low-level features like edges, texture and the basic shapes. Although these features are able to transfer to some extent, they are not sufficient for capturing clinically relevant patterns like tissue density variations, anatomical structures, and pathology-specific signals. This limitation is mainly evident in modalities like MRI and CT, in which the intensity distribution and spatial relationships are fundamentally different from natural images.

On the other hand, the domain-specific pre-training enables models to learn the feature representations which are directly aligned with medical imaging features. For example, 3D CNNs that are trained on volumetric CT data are able to capture inter-slice special continuity, while MRI-specific models can learn multi-sequence dependencies across T1, T2 and contrast-enhanced images. Similarly, the ultrasound-specific models become robust to speckle noise and operator variability when exposed to domain-relevant data distributions.

The effectiveness of these architectures depends strongly upon the alignment between the training data and the target clinical task. The case where the pre-training data closely matches the modality and task distribution, the models tend to perform well. This allows them to learn meaningful representations. Conversely, when there is a mismatch in the domain, the performance degrades which leads to poor generalisation, increased false positives and unreliable clinical predictions. Therefore, it is the combination of architecture design and domain-specific pre-training which determine the success of AI systems in medical imaging and not the architecture alone.

Foundations of Pre-Training in Medical Imaging¶

¶

The process of training an AI model on a large dataset to learn general feature representations before fine-tuning it on a specific task is referred to as “Pre-Training”. In medical imaging, general-purpose pre-training on foundational AI models (e.g. ImageNet) mostly provides an insubstantial benefit due to the fundamental differences between natural and medical images, which include modality specific intensity distributions and anatomical structures (Raghu et al., 2019).¶

¶

Domain-specific pre-training in AI models addresses this limitation by training the models directly on suitable medical datasets, thereby enhancing the learning of clinically applicable features and improving robustness to modality-specific noise and variability. This study builds on the framework to evaluate how pre-training strategies guide model reliability

across various imaging modalities.¶
¶

Methodology

This study is a narrative review that synthesizes existing literature on the role of domain-specific pre-training of AI models in medical imaging. Relevant studies were identified through the different searches conducted on databases such as PubMed, IEEE Xplore, and arXiv. using kKeywords which includedused were "Medical Imaging", "Transfer Learning", "Domain-Specific Pre-Training", "MRI Deep Learning", "CT Segmentation", "Ultrasound AI", and "Chest X-ray Deep Learning". Inclusion criteria included the peer-reviewed publications and some widely cited research papers that provided empirical evaluation of AI models in various medical imaging tasks. Studies which compared domain-specific pre-training with general-purpose pre-training were prioritized. The selection process prioritized peer-reviewed publications and widely cited research papers that provided empirical evaluation of AI models in medical imaging tasks such as image classification, detection, and object segmentation. Studies which compared domain-specific pre-training with general-purpose pre-training were included and discussed on the basis of modality-specific challenges in depth in this paper. Papers which lacked methodological clarity or empirical validation were excluded. Exclusion criteria included the studies which lacked empirical validation or had unclear methodologies. Studies which were published between 2015 and 2023 were prioritized to reflect recent developments in deep learning. This review is qualitative in nature and does not perform a formal meta-analysis; instead, performance metrics such as AUROC, Dice scores, sensitivity, and specificity are interpreted as mentioned in the original studies, with proper consideration of the dataset differences and experimental conditions.The selection process prioritized peer-reviewed publications and widely cited research papers that provided empirical evaluation of AI models in medical imaging tasks such as image classification, detection, and object segmentation. Studies which compared domain-specific pre-training with general-purpose pre-training were included and discussed on the basis of modality-specific challenges in depth in this paper. Papers which lacked methodological clarity or empirical validation were excluded. This review is qualitative in nature and does not perform a formal meta-analysis; instead, performance metrics such as AUROC, Dice scores, sensitivity, and specificity are interpreted as mentioned in the original studies, with proper consideration of the dataset differences and experimental conditions.

Results

The findings across the reviewed studies indicated that domain-specific pre-training consistently improved performance, robustness, and clinical applicability of the AI models across the various imaging modalities. Table 1 summarizes the key imaging modalities, associated AI tasks, and challenges.

Table 1: Overview of various Imaging Modalities

Modality	Key AI Tasks	Example Domain Specific Models	Challenges Addressed
Chest X-ray (CXR)	Pneumonia detection	CheXNet	Low resolution and overlapping structures
Computed Tomography (CT)	Lesion detection, cancer segmentation	3D U-Net, multi-view CNN	False positives, scanner variability
Magnetic Resonance Imaging (MRI)	Tumor segmentation, classification	nnU-Net, MRI-pretrained CNN	High soft tissue contrast and multi-sequence data
Ultrasound	Organ segmentation, echocardiography analysis	US-specific CNNs	Speckle noises and operator dependencies

(Performance comparisons across studies should be interpreted cautiously, as variations in datasets, evaluation protocols, and class distributions significantly influence reported AUROC and Dice scores; Abbreviations: CNN - Convolutional Neural Network; AUROC - Area Under the Receiver Operating Characteristic Curve)¶

Hardware workings of the various medical imaging modalities and their limitations

The chest x-ray imaging is primarily based on the photons passing through the tissues of various densities present. While this method is fast and cost-effective, there are some limitations to it as well, which is that the X-rays have many 2D projections with multiple overlapping anatomical structures, which limit the depth perception and also the sensitivity factors to subtle pathologies such as early COVID-19 infiltrates and other kinds of severe lung diseases.

Computed tomography or CT, is widely used currently in the healthcare industry. The CT scans reconstruct multiple X-ray projections into volumetric 3-dimensional representations using various advanced tomographic reconstruction algorithms and therefore it offers superior resolution and also shows a proper and a better density variation, but it also has a major limitation that is it exposes the patient to a higher amount of radiation doses which is extremely harmful.

Magnetic resonance imaging or MRI relies completely on the nuclear magnetic resonance principles to capture the tissue-specific contrast through variations in the relaxation timings (T1, T2) and the pulse sequences. MRI also provides an exceptionally soft tissue contrast but on the other hand it suffers from extremely high cost, long scan times and also scanner-dependent variability.

Ultrasound imaging, uses high frequency sound waves to generate real-time images. The advantage of this is that it is safe, portable and cheaper compared to the other imaging modalities but it is highly sensitive to speckle noises and operator dependency.
¶

¶
Chest X-Ray Imaging

One of the most widely used diagnostic modalities commonly used in clinical practices is the ~~c~~Chest X-ray (CXR). This is generally used for the detection and also monitoring of pulmonary and cardiovascular diseases. Since the cost is low, the outcome is prompt, and it is also broadly available, chest radiology is considered to be the primary and the first imaging tool for detecting conditions such as pneumonia, tuberculosis, lung cancer, COVID-19, pleural effusion, heart failure and other such things.

However, chest X-rays are not free from complex analytical challenges. Plenty of such challenges are present for automated systems and AI models. One of the important reasons for this is that since the images are two-dimensional (2D), the overlapping anatomical structures and subtle intensity variations require expert interpretation. X-ray analysis is therefore very complex, and various studies show that the AI models pre-trained on domain-specific data mostly surpass general-purpose foundational AI models in clinically relevant tasks when evaluated under comparable conditions (Raghu et al., 2019; Litjens et al., 2017).

This type of domain-specific pre-trained model becomes familiar with clinical terminologies, even the critical, advanced and complex ones. Not only this, but these models also become familiar with meaningful radiographic patterns, for example, lung-field asymmetries, interstitial markings and subtle opacities. These were generally missed or misinterpreted in the case of AI models that are pre-trained on natural images. General AI models that are pre-trained on natural images like ImageNet have difficulty perceiving subtle disease-specific patterns, due to negligible visibility to relevant medical imaging features (Zech et al., 2018).

One of the most prominent examples of such training in the case of chest X-ray imaging is *CheXNet,* proposed by *Rajpurkar et. al. (2017)*. Beyond pneumonia, various recent studies and literature have demonstrated the effectiveness of the usage of domain-specific AI models in detecting COVID-19 related things, tuberculosis cavitation and lung nodules as well, which are often subtle and easily missed by the general-purpose AI models. Since it was trained on large-scale expertly labelled X-ray datasets, CheXNet achieved state-of-the-art performance on multiple thoracic diseases, including pneumonia, with AUROC scores typically in the low-to-mid 0.80s on the ChestX-ray14 test set, comparable to radiologist-level performance (Rajpurkar et al., 2017).

¶

However, AUROC alone is sometimes not sufficient for cross study comparison as the datasets, disease prevalence, and evaluation protocols differ greatly. It was also found that CheXNet illustrated improved diagnostic performances in several reviewed parameters as compared to the general convolutional neural networks (CNNs). On the other hand, the performances of general-purpose CNNs pre-trained on natural images were stated to show limited generalizability in certain clinical settings (Zech et al., 2018). Zech et al. (2018) also showed that the AI models that were trained on chest X-rays from a single institution were unable to generalize to external datasets. Especially when pretrained on non-medical images.

From these findings, we can conclude that domain-specific pre-training will not only give more accurate results but will also reduce data bias to a large extent, therefore increasing robustness and reliability in real-world clinical environments. There are more recent studies that comparatively evaluated and found that the models pretrained on large-scale chest X-ray datasets consistently performed much better than general-purpose vision models of comparative size in both the classification and segmentation tasks (Irvin et al., 2019). The AI models that are pretrained on chest radiography are much aligned with radiological reasoning. Due to this the findings are better, and false positives are much less. This alignment is very important; since these are related to health, a small misclassification may lead to serious clinical consequences.

These findings indicate that extensive evidence has been confirmed in both foundational and recent studies that domain-specific pre-training is a significant factor on which AI performance in the case of chest X-ray imaging depends. These findings very clearly state that modality-aware pre-training is very important and essential for AI systems in chest radiography to be clinically trustworthy, and this provides a strong motivation to extend these domain-specific training strategies across other medical imaging modalities. A large chest radiography dataset contains 224,316 images from 65,240 patients (Irvin et al., 2019). The deep learning models which were trained on chest X-ray images specifically achieved a higher multi-pathology classification performance compared to the general Foundation AI models.

Table 2: Chest X-ray AI Model Comparison

Model	Pre-training	Task	Performance	Notes
CheXNet	Chest X-ray dataset	Pneumonia classification	AUROC ≈ 0.82–0.85	Evaluated on ChestX-ray14 dataset
CNN	ImageNet	Pneumonia classification	Performance varies across datasets; often lower generalization in clinical settings	Generalization issues reported (Zech et al., 2018)
CheXNet	Chest X-ray dataset	Multi-disease detection	Performance depends on dataset and labeling quality	Supported by multiple studies

(As shown in Table 2, domain-specific models such as CheXNet demonstrate improved robustness and generalizability compared to general-purpose CNNs, particularly under cross-institutional evaluation settings; Performance comparisons across studies should be interpreted cautiously, as variations in datasets, evaluation protocols, and class distributions significantly influence reported AUROC and Dice scores; Abbreviations: CNN - Convolutional Neural Network; AUROC - Area Under the Receiver Operating Characteristic Curve).¶

Computed Tomography (CT) Imaging

Computed tomography imaging plays a vital role in day-to-day clinical workflows, especially in oncology and trauma assessment tests. CT scan imaging is also widely used in the detection of pulmonary embolism, stroke tests, intracranial hemorrhage identification and COVID-19 tests as well. CT scans provide high-resolution three-dimensional (3D) volumetric data which capture very finely grained anatomical structures and details across a wide range of slices. These characteristics demand AI models which are capable of properly understanding the complex 3D spatial relationships,

tissue density variations, and subtle pathological patterns.

Computed tTomography (CT) imaging represents one of the most complex modalities for reliable automated analysis. General-purpose foundational AI vision models which are trained on two-dimensional natural images, most of the time¶
¶

struggle to interpret the volumetric CT data effectively. These models lack intrinsic mechanisms which are necessary for learning the spatial continuity across various slices and may sometimes fail to capture clinically relevant 3D contextual data.

In contrast, the AI models trained on volumetric CT data demonstrate highly improved sensitivity and robustness in the clinical detection and segmentation tasks (Çiçek et al., 2016; Setio et al., 2017). Previous studies demonstrate that CT-specific models which are trained on volumetric data improve cancer detection sensitivity and lesion localization to a great extent when compared to 2D vision models pretrained on natural images (Çiçek et al., 2016; Setio et al., 2017). Several studies show high improvements in sensitivity for lesion detection when CT-specific 3D architectures are used in comparison to 2D vision models (Setio et al., 2017).

CT-specific domain-trained models have been demonstrated to reduce false positives in lesion detection tasks, which is an extremely important and crucial factor in minimizing unnecessary follow-up procedures (Setio et al., 2017). These findings show the disadvantages of the general-purpose foundational AI models in handling volumetric medical data and also highlight the necessity of modality-aware pre-training.

Another major advantage of CT-specific domain pre-training is in its ability to account for scanner-related variability and reconstruction artefacts. The CT images vary a lot, depending a lot on the scanner manufacturers, the radiation dose protocols and the reconstruction algorithms. The pre-training of AI models with specific domain data enables the AI models to learn invariant features properly and in depth so that they can generalize across such variations, whereas general-purpose vision models often wrongly interpret reconstruction artefacts as pathological features and hence limit their clinical reliability (Litjens et al., 2017; Roberts et al., 2021).

Overall, this research and these studies demonstrate that CT imaging shows fundamental shortcomings in general-purpose foundational vision AI models, and it also strongly supports the adoption of domain-specific data pre-training methods. For complex and critical applications like cancer detection and trauma diagnosis, the domain-aware CT AI models are extremely needed for achieving a clinically reliable and trustworthy AI performance.
¶

Table 3: CT AI Performance

Model	Pre-training Type	Task	Performance	Notes
3D CNN	CT-specific dataset	Lesion detection	Higher sensitivity reported in controlled studies	Captures volumetric context
2D CNN	ImageNet	Lesion detection	Lower sensitivity in volumetric tasks	Limited 3D understanding

(As summarized in Table 3, CT-specific models such as 3D CNNs outperform 2D models due to their ability to capture volumetric spatial relationships; Performance comparisons across studies should be interpreted cautiously, as variations in datasets, evaluation protocols, and class distributions significantly influence reported AUROC and Dice scores; Abbreviations: CNN - Convolutional Neural Network; AUROC - Area Under the Receiver Operating Characteristic Curve).¶

Magnetic Resonance Imaging (MRI)

Magnetic rResonance iImaging has emerged as one of the most important imaging tools in modern-day medical diagnosis, especially in the fields of neurological imaging, oncology, and musculoskeletal imaging. Unlike other medical imaging tools such as X-ray technology-based medical imaging modalities, the MRI image has very high-resolution volumetric data with very subtle contrast in soft tissues that are demonstrated through appropriate image acquisition protocols such as the T1-weighted image, the T2-weighted image, and the contrast-acquired image (T1ce image). This very multi-sequence and high-dimensional nature of the MRI data makes it in particular less acceptable for direct transfer learning from the natural image datasets.

The analysis of an MRI image becomes more complex for a general AI system, this is because there is a large variation in the intensities specified to MRI image capturing tools which is again combined with a large amount of anatomical details

which were absent in natural images. Due to this, MRI serves as a critical platform for evaluation of the effectiveness of¶
¶

domain-specific pre-training in medical imaging models. The transfer learning from the natural image datasets most of the times provide a slight benefit in capturing the modality-specific intensity variations which are inherent to MRI data (Raghu et al., 2019).

A growing body of the existing literature continuously shows that the pre-trained AI models, those that are pre-trained on specific MRI datasets are much better than the general-purpose foundational AI models of same sizes across a wide range of tasks including tumor detection, classification and segmentation. Raghu et al. (2019) elaborately investigated the effectiveness of features learnt from the datasets which contained natural images in medical imaging tasks and demonstrated that the ImageNet pre-training provided very limited benefits for applications which were based on magnetic resonance imaging.

Their findings showed that when the AI models of the similar sizes were compared, the AI models which were pretrained directly on the MRI datasets consistently outperformed the ImageNet pretrained counterparts across medical tasks such as ~~tumour~~tumor detection and ~~tumour~~tumor segmentation. This limitation comes out most of the time because the MRI data contains a lot of modality-specific complex features and details, such as tissue contrast mechanisms and anatomical structures, which are definitely not present in the natural image datasets. Therefore, the general-purpose foundational AI models find it very difficult and complex to learn these complicated and meaningful MRI details and representations without proper domain exposures.

Studies support these findings by demonstrating that the AI models trained directly on MRI data achieve a greater and a better tumor segmentation performance across multiple sequences which include T1, T2, and also the contrast-enhanced MRI when compared to the AI models which are pretrained on natural images (Isensee et al., 2021; Bakas et al., 2018). This advantage in the performance has been continuously observed on the public brain tumor segmentation benchmarks such as BRATS.

Across the various public brain tumor segmentation benchmarks, MRI-specific training strategies, such as nnU-Net, have been reported to achieve Dice similarity coefficients around or above 0.80 on the BRATS benchmarks, depending on the task and validation split (Isensee et al., 2021), while the AI models relying mainly on ImageNet pre-training consistently show a lower segmentation accuracy (Raghu et al., 2019; Isensee et al., 2021). This throws light on the limitations of the ImageNet based pre-training for the MRI-specific segmentation tasks.

These differences in the performances are particularly significant in clinical matters, where proper and precise tumor boundary demarcation directly influences the treatment planning and the diagnostic outcomes. In addition to these segmentation tasks, the MRI-specific foundational AI models have also demonstrated a much better performance in classification tasks like ~~tumour~~tumor grading and the detection of diseases. These tasks include the stroke outcome prediction as well.

Multiple studies and literature have reported that the MRI-specific representational learning improves the disease classification and also the tumor grading performances when compared to the general-purpose foundational vision AI models, especially when the model sizes are the same (Raghu et al., 2019; Litjens et al., 2017). These improvements show the importance of AI models learning MRI-specific anatomical priors, which the general vision AI models often fail to capture effectively and properly. Other important and crucial factors influencing the MRI performances are scanner variability and acquisition heterogeneity. MRI images vary greatly across the scanner manufacturers, field strengths and patient demographics. Domain-specific pre-training with proper and wide MRI datasets exposes the AI models to the scanner and acquisition variability during learning, which leads to improved robustness and cross-institutional generalization, in contrast, the general-purpose models often fail to adapt reliably and effectively to the unseen MRI distributions (Raghu et al., 2019; Roberts et al., 2021). Such robustness is extremely crucial for the safety and the reliability factor of clinical deployment of the MRI based AI models.

Overall, the extensive empirical evidence confirms that the MRI-based tasks substantially benefit from domain-specific pre-training with MRI-specific data when the AI models of the similar sizes are compared. These findings properly highlight that the MRI's modality-specific complexity requires proper and appropriate learning strategies and that the reliable clinical deployment of AI in MRI cannot rely solely on the general-purpose vision pre-training.

¶

Table 4: MRI AI Performance

Model	Pre-training Type	Task	Dice Score / AUROC	Notes
nnU-Net	MRI dataset	Brain tumor segmentation	≈ 0.80+	BRATS benchmark (Isensee et al., 2021)
CNN	ImageNet	Brain tumor segmentation	Typically lower than MRI-specific models	Limited transferability

(Table 4 highlights that MRI-specific models outperform ImageNet-pretrained models due to their ability to learn modality-specific tissue contrast and multi-sequence dependencies; Performance comparisons across studies should be interpreted cautiously, as variations in datasets, evaluation protocols, and class distributions significantly influence reported AUROC and Dice scores; Abbreviations: CNN - Convolutional Neural Network; AUROC - Area Under the Receiver Operating Characteristic Curve).¶

Ultrasound Imaging

Ultrasound imaging, due to its non-invasiveness, real-time acquisition and low cost, is being widely adopted across a wide range of clinical domains (e.g., abdominal imaging, obstetrics and cardiology). Other important applications include vascular assessment and echocardiography assessments as well. However, ultrasound imaging carries a number of complex and unique challenges, e.g., low signal-to-noise ratio, speckle noises, operator dependency, and considerable variability in image qualities. The general foundational vision AI models which are trained on natural image datasets may misinterpret speckle noises and acquisition artefacts as meaningful features, thereby leading to unreliable predictions (Xie et al., 2018; Chen et al., 2019).

On the other hand, domain-specific pre-training on large-scale ultrasound data enables AI models to learn robust clinically relevant patterns despite the substantial noise and operator variability (Xie et al., 2018; Chen et al., 2019). Multiple surveys in the literature have shown that deep learning models pre-trained on domain-specific ultrasound data consistently yield strong performance compared to general foundational AI models with comparable parameter sizes (Xie et al., 2018; Chen et al., 2019).

The parameters such as accuracy, f1 score and robustness under the noise consistently favour the domain trained AI models. Specifically, in the noisy and operator-dependent imaging condition, AI models pre-trained on domain-specific pre-training data consistently outperform the general-purpose foundational AI models of comparable sizes (Xie et al., 2018; Chen et al., 2019).

These differences are very critical when AI is taken and applied under real-world clinical settings, especially when ultrasound images vary significantly across devices and patient populations. This significant performance gap between general-purpose foundational vision AI models and ultrasound-specific AI models accentuates the need for modality-aware training in AI models in the healthcare field for trustworthy, accurate, and reliable results.

¶

Table 5: Ultrasound AI Performance

Model	Pre-training Type	Task	Performance	Notes
CNN	Ultrasound dataset	Clinical tasks	Improved robustness under noise conditions	Domain adaptation beneficial
CNN	ImageNet	Clinical tasks	Performance sensitive to noise and variability	Less robust

(As shown in Table 5, domain-specific ultrasound models demonstrate improved robustness to noise and operator variability compared to general-purpose models; Performance comparisons across studies should be interpreted cautiously, as

variations in datasets, evaluation protocols, and class distributions significantly influence reported AUROC and Dice scores; Abbreviations: CNN - Convolutional Neural Network; AUROC - Area Under the Receiver Operating Characteristic Curve).¶

Discussion

Through the broad spectrum of the imaging modalities discussed in this review paper, one could notice a similar trend suggesting that AI models which are pre-trained on domain-specific medical data, often tend to be more powerful and reliable when contrasted with the foundational AI models. Various findings from medical imaging studies suggest that modality specific details and pathologies are most of the times absent in the natural image datasets due to which they may result in diminished performance in certain clinical tasks. Hence, we may conclude that the general-purpose foundational AI models rely on dataset-specific patterns and shortcuts to a great extent in generalizing the practical applications in the healthcare sector (Tajbakhsh et al., 2016; Raghu et al., 2019; Zech et al., 2018; Oakden-Rayner, 2020; Roberts et al., 2021). Sometimes, the general vision AI models may find it complex in terms of variations in image acquisition conditions, image reconstruction algorithms, and patient demographics as well. These variances are mostly better handled through domain-specific pre-trainings in AI models due to which more reliable interpretations are developed.

The domain-specific image datasets have also shown great future scope to improve. There are a few largely used image datasets in the domain-specific area which are not fully certified or verified in a correct manner by radiologists. Hence, residual variances may remain (Rajpurkar et al., 2017). However, a few image datasets lack a variety of demographic populations due to which AI-related systems dedicated to healthcare may raise concerns in terms of fairness and reliability (Oakden-Rayner, 2020; Roberts et al., 2021). To overcome these constraints, more interactions are needed between humans, doctors and AI researchers. Domain-specific AI models are not intended to replace doctors or radiologists. Instead, they serve as efficient tools to enhance human-AI collaboration. When these models are correctly developed with accurate training and used practically, they can act as a reliable second reader. This can help relieve radiologists from heavy workloads and aid in sound clinical decision-making. However, using radiologist-verified datasets is crucial for improving fairness, safety, and reliability.¶

¶

The well-trained medical AI models can support radiologists by significantly reducing their workload and also acting as reliable readers. To understand the environments under which these AI models give clinically safe and trustworthy diagnostics results is extremely necessary for the application of AI in real world clinical practices.

Though there are substantial advantages of domain-specific pre-training, there are certain limitations as well. One major challenge is the limited availability of large-scale, high-quality annotated medical datasets, because expert annotation is both time-consuming and also very expensive. Moreover, many existing datasets lacks sufficient demographic diversity, which may lead to bias and limit the generalizability of AI models across different populations (Roberts et al., 2021; Oakden-Rayner, 2020). Furthermore, variability in imaging protocols, scanner types, and acquisition settings across institutions can affect model robustness despite domain-specific pre-training. The computational costs related to training large-scale domain-specific models also remain high, thereby posing practical constraints for widespread adoption.

Comparing with alternative approaches like transfer learning and emerging techniques like meta-learning, the domain-specific pre-training offers improved feature alignment, but may lack flexibility when they are applied to unseen tasks or modalities. Transfer learning enables faster adaptation with limited data, while meta-learning focuses on rapid generalization across tasks. However, these approaches may not capture modality-specific characteristics as efficiently as the domain specific pre-training. Future research should therefore focus on and explore hybrid strategies that combine domain-specific learning with adaptable frameworks, this will improve scalability and generalization.

~~Summing up this review, domain specific pre-training is not merely a decision-making technique, but also a basic requirement for the proper development of trustworthy AI models in the fields of medical practices especially medical imaging. By incorporating the findings of the various imaging modalities, this paper not only provides a comprehensive study of the existing literature, but also gives an insight for the future development process of strong, dependable and clinically channelized AI systems.¶~~

¶

~~Limitations of Domain-Specific Pre-Training¶~~

~~Although domain-specific pre-training shows some promising advantages, some limitations have been introduced, which can be related to data availability, annotation costs, and legal issues. In addition, medical datasets need high-quality annotation, which is costly, and datasets may have limited demographic variation, which can affect fairness (Roberts et al., 2021). Finally, computational costs of large-scale domain-specific training are still high.¶~~

¶

Conclusion

This review clearly demonstrates that the need for domain-specific pre-training is a significant necessity for the reliable AI system for medical images. Future research should focus on multi-modal data integration, cross-institutional validation, improved dataset diversity, and real-world clinical deployment of AI systems. The trained AI systems for modality-aware datasets have demonstrated a reliable and phenotypically accurate outputs than the foundational vision AI systems for general applications. The outcomes also represent the significance of dependability for medical applications. The dependence on general vision AI systems provides a clearer understanding of the constraints of general vision AI systems for healthcare applications. The application of domain-specific Pre-trained AI systems would provide improved human-AI collaboration to align human behavior with better outcomes. Further research work would be required in multi-modal integration for successful utilization of AI systems for¶

¶

healthcare applications. Finally, the development of precise and reliable AI systems would be required for healthcare applications to incorporate understanding of the domain. Future work should also focus on systematic comparisons between domain-specific pre-training, transfer learning, and meta-learning frameworks so as to better understand their relative strengths and their limitations in the real-world clinical applications.

¶
¶

Acknowledgement

The author of this paper acknowledges the contributions of all the researchers whose work formed the foundation of this review paper and also extends his sincere gratitude to his mentors and peers for their valuable guidance and feedback all throughout this research process.

¶

Bibliography

1. 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.
2. Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems, 32.
3. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225.
4. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine, 15(11), e1002683.
5. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019, July). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 590-597).
6. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods, 18(2), 203-211.
7. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., ... & Jambawalikar, S. R. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629.
8. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning?. IEEE transactions on medical imaging, 35(5), 1299-1312.
9. Setio, A. A. A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., Van Riel, S. J., ... & Van Ginneken, B. (2016). Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE transactions on medical imaging, 35(5), 1160-1169.
10. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016, October). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In International conference on medical image computing and computer-assisted intervention (pp. 424-432). Cham: Springer International Publishing.
11. Roth, H. R., Lu, L., Liu, J., Yao, J., Seff, A., Cherry, K., ... & Summers, R. M. (2015). Improving computer-aided detection using convolutional neural networks and random view aggregation. IEEE transactions on medical imaging, 35(5), 1170-1181.
12. Smistad, E., Falch, T. L., Bozorgi, M., Elster, A. C., & Lindseth, F. (2015). Medical image segmentation on GPUs–A comprehensive review. Medical image analysis, 20(1), 1-18.
13. Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S. X., ... & Wang, T. (2019). Deep learning in medical ultrasound analysis: a review. Engineering, 5(2), 261-275.
14. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC medicine, 17(1), 195.
15. Oakden-Rayner, L. (2020). Exploring large-scale public medical image datasets. Academic radiology, 27(1), 106-112.
16. Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., ... & Schönlieb, C. B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence, 3(3), 199-217.
17. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Cham: Springer.

Response to Reviewer 1

I sincerely thank the reviewer for the thoughtful and constructive feedback provided throughout the review process. I deeply appreciate the reviewer's continued guidance regarding the technical depth, architectural discussion, critical analysis, and future directions in the manuscript. These comments significantly helped strengthen the scholarly rigor and analytical quality of the paper. I have carefully revised the manuscript in accordance with all recommendations and highlighted all newly added or modified sections in bold within the revised manuscript.

Substantive Comments

Comment 1: I thank you for more carefully refocusing the title and Intro to emphasize your exact application instead of a generic discussion of pretraining.

Response 1: Thank you so much for so an encouraging and helpful feedback, and continued guidance as well. I greatly appreciate the reviewer's patience and observations regarding the framing and scope of the manuscript. The revised Introduction was further refined properly to maintain a stronger focus on the role of domain-specific pre-training in medical imaging rather than just discussing pre-training in a broader generic context.

Comment 2: The most important edit I would like to re-emphasize from my original comments is that you should thoroughly describe the machine learning architectures used in your Foundations of Pre-Training section. This should emphasize a combination of images and equations, not text, since architectures can't be explained precisely in words. Please show how the training and pre-training processes work for, at least, a few example cases you describe in the literature.

Response 2: Thank you so much for this highly valuable recommendation. I have substantially expanded the section titled "Model architectures and the pre-training mechanisms in medical imaging" to provide a deeper discussion regarding - CNN and U-Net architectures, hierarchical feature extraction, optimization behaviour, cross-entropy loss minimization, latent feature representations, transferability of learned features and modality-specific representation learning as well. I additionally incorporated more detailed discussion explaining how pre-training modifies feature representations and optimization trajectories across different imaging modalities. The existing CNN and U-Net figures were properly retained and the accompanying architectural explanations were significantly strengthened to improve technical clarity while maintaining consistency with the narrative review format of the paper.

Comment 3: I would still encourage you to more carefully state what makes the architectures work and not work in different situations. They don't always work. Why? When they do work, why? That leads to a critical understanding of the field.

Response 3: Thank you so much for this valuable suggestion. I have carefully revised multiple modality sections and the Discussion section to include deeper critical analysis regarding - feature transferability, dataset bias, optimization limitations, shortcut learning, institutional overfitting, volumetric contextual learning and external generalization failures. I have additionally discussed how the effectiveness of pre-training strategies depend strongly upon - modality complexity, task difficulty, dataset scale and feature similarity between source and target domains. This helped me in improving the manuscript's analytical depth beyond descriptive literature summarization.

Comment 4: The limitations of the domain-specific pretraining section should either be longer or be integrated into the Discussions section; it seems orphaned right now.

Response 4: Thank you so much for this important and valuable suggestion. In accordance with the recommendation, I have properly integrated the discussion of limitations directly into the broader Discussion section rather than maintaining it as an isolated subsection. This improved the overall flow, coherence, and critical continuity of the manuscript. The revised Discussion section now includes expanded analysis regarding - dataset limitations, demographic bias, scanner variability, computational cost and external validation challenges.

Comment 5: You should discuss more of the challenges and opportunities for these methods looking forward. What are competing methods (like meta-learning and transfer learning)? How do they perform? How might these competing approaches perform in the future?

Response 5: Thank you so much for this valuable suggestion. I have significantly expanded the forward-looking analysis of the manuscript by - comparing domain-specific pre-training with transfer learning, meta-learning, and self-supervised learning approaches, discussing tradeoffs between these competing methodologies, analyzing their respective strengths and limitations and introducing a new dedicated "Future Directions" section. This newly added section discusses - multimodal learning,

explainable AI, uncertainty estimation, benchmarking frameworks, self-supervised learning and prospective clinical deployment challenges. These additions helped me provide a more actionable and literature-grounded future research directions.

Response to Reviewer 2

I sincerely thank the reviewer for the thoughtful and constructive feedback provided throughout the review process. I greatly appreciate the reviewer's positive evaluation of the manuscript's structure and clarity, as well as the valuable recommendations regarding methodology transparency, comparative analysis, writing quality, and future directions. These comments substantially helped strengthen the academic rigor and analytical depth of the manuscript. I carefully revised the manuscript in accordance with all recommendations and highlighted all newly added or modified sections in bold within the revised manuscript.

Substantive Comments

Comment 1: Thank you for the corrections you have submitted. I really appreciate the clarity and structure of your manuscript. The way you present and organise your discussion works well, and it reflects a strong grasp of the subject area. Your synthesis of the literature is clear and accessible, making the paper engaging to read. I particularly like how you guide the reader through each section in a logical and coherent way.

Response 1: Thank you very much for these encouraging words regarding the organization and presentation of the manuscript. I sincerely appreciate the reviewer's appreciation of the manuscript structure and literature synthesis and I am deeply grateful for the detailed feedback that helped me in further improving the paper.

Comment 2: That said, there are a few areas where some refinement would further strengthen the paper. The methodology section needs to explain your approach in more detail. While you identify the work as a narrative review, it would be beneficial to outline your search strategy more explicitly, including the databases used, inclusion and exclusion criteria, and any timeframes applied. This would improve transparency and enhance the academic rigor of the paper.

Response 2: Thank you so much for this valuable recommendation. I have substantially expanded the Methodology section to improve transparency and reproducibility of the review process. Specifically, I have clarified - the databases searched, the keywords used, inclusion and exclusion criteria, prioritization strategy for highly cited studies, full-text review procedures and the criteria used to assess methodological clarity and empirical validation. I have also additionally clarified that the literature screening and relevance assessment were conducted manually through full-text review by the author.

Comment 3: You might also consider deepening the critical engagement with the literature. It might be also good to section your paper with a Literature Review chapter to focus your flow of thought. While the summaries are clear, it would add value to compare studies more directly, particularly in terms of datasets, evaluation approaches, and limitations. It might be worth adding a more consistent way of presenting performance metrics such as AUROC and Dice scores to allow clearer cross-study comparison.

Response 3: Thank you so much for this insightful suggestion. I revised multiple modality sections and the Discussion section to include stronger cross-study comparisons and deeper critical analysis regarding - dataset differences, external validation, task complexity, modality-specific limitations, feature transferability and robustness tradeoffs. I have additionally contextualized performance metrics such as AUROC and Dice scores more carefully to avoid direct comparisons across substantially different datasets and evaluation settings.

Comment 4: In terms of writing, I really like the clarity in several sections; however, there are areas where the phrasing becomes repetitive or slightly informal. A focused language edit would help improve conciseness and maintain a consistent academic tone throughout. It might be also good to go through your language and de-AI is so that the flow of the content is clear and deep.

Response 4: Thank you for this important observation and feedback. I have carefully revised the manuscript to improve conciseness, consistency and academic tone throughout the paper. Specifically, I have reduced repetitive phrasing and replaced several broad or inflated statements with more mechanism-focused and analytical discussions. Repeated statements regarding domain-specific models and general-purpose models were carefully revised to include more nuanced comparative reasoning and modality-specific distinctions as well.

Comment 5: It might also be worth addressing the formatting issue in the ultrasound section, where part of the text is obscured by an overlapping image, as this affects readability.

Response 5: Thank you for identifying this formatting issue. I have corrected the formatting and layout problems in the ultrasound section and carefully reviewed the formatting consistency throughout the manuscript to improve readability.

Comment 6: Finally, you might consider expanding the conclusion to include more explicit and actionable future research directions, particularly around real-world implementation and data-related challenges.

Response 6: Thank you for this valuable suggestion. I have now carefully expanded the future-oriented discussion substantially

and also created a new dedicated “Future Directions” section before the Conclusion. This section now includes discussion regarding - multimodal learning, explainable AI, external validation, self-supervised learning, uncertainty estimation, benchmarking frameworks and real-world clinical deployment challenges. These additions were also supported using themes and limitations identified throughout the reviewed literature.

The author has done a good job in revising the paper to better align with the expectations of an academic publication, as suggested by the Reviewers. However, **some issues remain that need to be rectified before this paper can be considered for acceptance for publication.** These include:

Introduction:

- There are now 2 aims towards the end of the Introduction (last and fourth-last paragraphs). Please consolidate these to only have one set of aims in the last paragraph.

Methodology:

- How were "some widely cited research papers that provided empirical evaluation of AI models in various medical imaging tasks" found? Was this by searching the reference lists of included papers, or were they searched in the databases sorted by citation numbers? This needs more clarification.
- Who conducted the search? How were papers deemed relevant or not? This needs to be clearly articulated to show whether these decisions were made by one person or through some form of consensus.
- "Studies which compared domain-specific pre-training with general-purpose pre-training were prioritized. The selection process prioritized peer-reviewed publications and widely cited research papers that provided empirical evaluation of AI models in medical imaging tasks such as image classification, detection, and object segmentation." What were these papers prioritised over?
- What criteria were used to determine that a paper was to be excluded based on "Papers which lacked methodological clarity or empirical validation were excluded"? Who made these decisions, and what would constitute a lack of methodical clarity or empirical validation?
- Were full-text versions of papers reviewed, or was it just the Title and Abstract?

Results:

- Tables 2-5 are still not mentioned in-text like Table 1 is.
- Please check abbreviations included under each table actually refer to the Table's contents. e.g., AUROC is included in Table 1, but this term is not even used in the table? Further, nnU, CheXNet, and US are used in Table 1 and are not defined. This is an issue across all tables.

Overall, <u>when writing a response to reviewers, it is best practice to explicitly write in the response to each question exactly what change was made, rather than just stating that a change was made. Moreover, if any changes were made to the manuscript, the relevant excerpt from the manuscript should be placed under the author's response</u> (or, at the very least, inclusion of line numbers where the change can be found). This allows Reviewers to check that the change has been made appropriately, especially if many changes were requested. This would also show a better understanding and resolution of the comments. Hence, **I am also requesting that the author redo their response letters from the last round and make sure that all relevant changes are explicitly referred to in these response letters.**

Decision: **Accept, *conditional* on minor revisions.**

Review: I appreciate the evident amount of effort that the author put into their revisions, and it is clear that they have partially and in some cases fully addressed the reviewers' concerns. However, reading through the revisions in more detail, I am concerned that the author did not fully provide quite the level of detail requested for the search criteria and architecture explanation. The technical depth and reproducibility of the review process still lack depth, although at least far better than before, when they basically had nothing described.

Hence, my remaining concerns largely pertain to concerns already brought up by Reviewer 1, but also some from Reviewer 2.

In particular:

- (R1) There is still very little detail about how pretraining actually modifies representations or optimization, and I would appreciate a bit more explanation about the mathematics or model architecture.
- (R1) CNN/U-Net figures were added, but there are still no equations, training pipeline diagrams, feature-learning analysis, loss-function discussion, nor explanation of transfer mechanisms.
- (R1) I do appreciate that the paper now gives some qualitative reasons for why models tend to succeed vs. fail, but can you do more? E.g. you can talk about feature transferability, data regime dependence, specific failure modes, model misrepresentation, etc. (maybe just pick a few of these to add)
- (R1) I still feel like the paper reads very summary-like rather than adding truly novel, critical analysis; a review paper goes far past just summarizing techniques and literature. Can you do some more direct comparisons between studies? Are there any limitations, conflicts, or contradictions in the literature? Can you compare and contrast some more learning techniques and competing approaches?
- (R2) The performance metrics are now contextualized better, but studies are still mostly discussed sequentially rather than critically compared. There is a need for deeper cross-study analysis.
- (R2) You added some more future work to the conclusion, but could you make it a bit more actionable? Ideally, it should be in a separate section and well-developed by hearkening to the literature and showing *where* it wants to point, and you should be justifying *why* these are epistemically valid and actionable future directions.
- (R2) Some sections could be less repetitive and replace that repetitive content with deeper, critical analysis. Please carefully par down on the redundant, inflated wording (e.g., "critical, advanced and complex ones," "reliable and trustworthy AI performance," "proper and appropriate learning strategies," etc.); more fluff and unnecessary verbosity are usually not good and prevent the reader from gleaning your more core points.
    - What do you mean by, "aligned with radiological reasoning?"

Moreover, following off the last point, I feel like the paper sort of repeats something along the lines of "modality-specific models are good + general-purpose models are bad => therefore,

domain-specific pretraining is important” across several sections. Even the cadence and vocabulary tend to be very similar. You should be doing more, such as comparing mechanisms, discussing exceptions, analyzing tradeoffs between models, and distinguishing between different modalities more carefully.

Enhancing the clinical reliability of AI models in medical imaging through domain-specific pretraining

Background: Artificial Intelligence (AI) has been integrated into many sectors, leading to improvements in accuracy and efficiency. AI has demonstrated improvements in image classification, object detection, and image segmentation tasks in which they can outperform conventional approaches under controlled conditions. Healthcare is one of the most critical domains because even minor diagnostic errors can significantly affect patient safety and treatment planning. Therefore, reliability is one of the most important factors for any AI system operating in the healthcare sector. Despite this, many AI models, which are primarily pre-trained on natural images, are used in various healthcare applications, which may limit their clinical reliability and generalizability.

Methods: This manuscript presents a qualitative narrative review of existing literature, comparing general-purpose pre-training strategies and domain-specific pre-training strategies across four major imaging modalities: chest X-rays, magnetic resonance imaging, computed tomography, and ultrasound. Relevant studies were identified through searches across major databases, including PubMed, IEEE Xplore and arXiv.

Results: The reviewed studies generally suggest that AI models pre-trained on domain-specific data demonstrate an improved performance, robustness, and clinical applicability across multiple medical imaging modalities when compared with AI models that are pre-trained on natural image datasets.

Conclusion: The reviewed evidence suggests that domain-specific pre-training represents a promising strategy for improving the clinical reliability and generalizability of AI systems used in medical imaging. Future research should focus on multimodal data integration, cross-institutional validation, real-world clinical testing, and systematic comparisons with alternative learning paradigms such as transfer learning and meta-learning.

Keywords: artificial intelligence, medical imaging, domain-specific pre-training, chest x-ray, magnetic resonance imaging, computed tomography, ultrasound, human-AI collaboration

Introduction

Medical imaging is vital for modern clinical diagnosis, allowing for non-disruptive imaging of anatomical structures, cells and tissues, required for treatment, prognosis, and disease mapping. Imaging modalities such as chest x-ray (CXR), computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound are widely used in the healthcare sector and are a critical component of the diagnostic workflow.

In recent years, AI, particularly deep learning (DL), has been a significant tool for medical image analysis, allowing noticeable improvements in diagnostic accuracy, efficiency, and clinical progress (Litjens et al., 2017). The success of DL in general computer vision, which includes tasks such as image classification, object detection, and object recognition using natural-image datasets, has accelerated the adoption of general-purpose foundational AI models in medical imaging. These AI models are mostly pre-trained on large-scale natural image datasets. Large-scale natural-image datasets include ImageNet, which is the most widely used dataset for pre-training DL models in computer vision. These AI models in medical imaging have shown promising accomplishment across a wide range of general visual tasks. As medical images are modality-specific, they differ greatly from natural images and thus require trained and expert clinical data interpretation. Consequently, AI models trained on non-medical data may not be generalized optimally for medical imaging tasks, potentially limiting their clinical reliability and robustness (Kelly et al., 2019; Roberts et al., 2021). Previous studies suggest that models pretrained on natural-image datasets may not adequately capture clinically relevant anatomical and pathological features, which can reduce performance on certain medical imaging tasks (Tajbakhsh et al., 2016; Litjens et al., 2017).

All imaging modalities face challenges and difficulties and thus demands domain-aware learning. MRI offers superior soft-tissue contrast through multiple imaging sequences, thereby enabling comprehensive assessment of neurological and musculoskeletal abnormalities, but requires an understanding of complex tissue-specific signal intensity variations. CT provides high-resolution volumetric imaging for applications such as oncology and trauma assessment. Meanwhile, ultrasound presents challenges such as speckle noise, operator dependency, and variable image quality. However, general foundational AI models are still used for many of these imaging tasks despite these modality-specific challenges.

The process of training an AI model on a large dataset to learn general feature representations before fine-tuning it on a specific task is referred to as pretraining. In medical imaging, general-purpose pretraining using natural-image datasets such as ImageNet often provides limited benefits due to the fundamental differences between natural and medical images (Raghu et al., 2019). The limited transferability of natural-image pretraining has been associated with differences in feature distributions and optimization behavior between medical imaging datasets and non-medical datasets. In natural-image datasets such as ImageNet, convolutional neural networks (CNNs) primarily learn edges, textures, colors, and object-level semantic features. However, medical imaging tasks oftentimes depend upon subtle grayscale intensity variations, tissue morphology and pathology-specific structures that differ greatly from natural images (Raghu et al., 2019; Tajbakhsh et al., 2016). As a result, pretrained representations derived from natural image datasets may converge towards clinically irrelevant features, thereby resulting in weaker generalization under cross-institutional or modality-shifted conditions. Domain-specific pretraining addresses these limitations by exposing AI models to medically relevant anatomical and pathological patterns during representation learning, thereby improving robustness to modality-specific variability. This review examines how different pretraining strategies influence the performance, robustness, and clinical reliability of AI models across various medical imaging modalities.

This paper aims to evaluate how domain-specific pre-training influences the clinical reliability, robustness and generalizability of AI models across four major medical imaging modalities: CXR, CT, MRI, and ultrasound. Furthermore, it critically compares domain-specific and general-purpose pre-training strategies, reviews the mechanisms underlying their performance differences and hence evaluates their implications for the development of clinically reliable AI system in medical imaging.

Model architectures and the pre-training mechanisms in medical imaging

In the context of this narrative review, mathematical or architectural derivation in detail is beyond the scope. The key concept is to properly understand how different model architectures interact with pre-training strategies and how this interaction influences the performance across medical imaging tasks. As far as capturing spatial hierarchies, contextual information, and modality-specific patterns, these architectures differ.

Figure 1 illustrates a clear conceptual distinction between general-purpose and domain-specific pre-training strategies. While the general-purpose models transfer representations learned from the natural image datasets, domain-specific approaches learn modality relevant features (anatomical and pathological) directly from the medical imaging data, which improves both robustness and clinical reliability to a great extent.

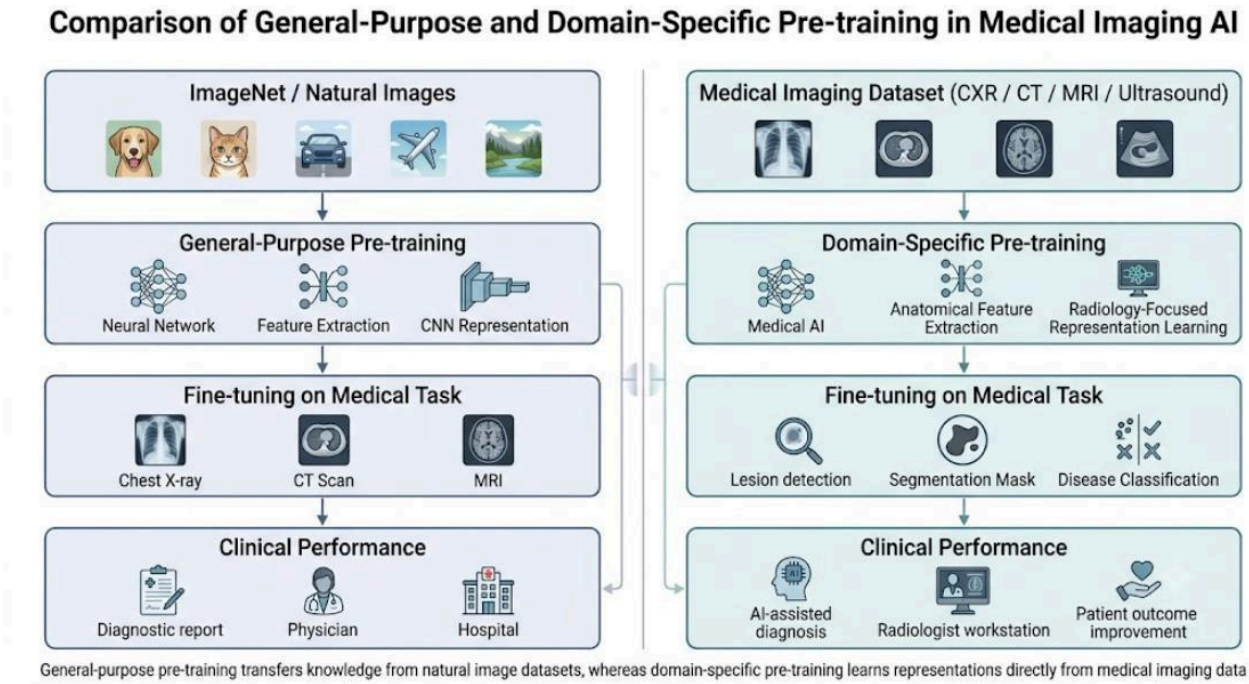


Figure 1. *Conceptual comparison of general-purpose and domain-specific pre-training pipelines in medical imaging. General-purpose pre-training transfers representations which are learned from the natural image datasets whereas the domain-specific pre-training learns representations directly from the medical imaging data before task-specific fine-tuning.*

Encoder-decoder architectures such as U-Net have become very important for medical image segmentation, which involves identifying and precisely delineating anatomical structures or pathological regions within medical images. This is because they combine contextual feature extraction with precise spatial localization. As shown in Figure 2, skip

connections preserve fine-grained anatomical information while integrating higher-level semantic features, thereby making U-Net highly effective for biomedical segmentation tasks (Ronneberger et al., 2015).

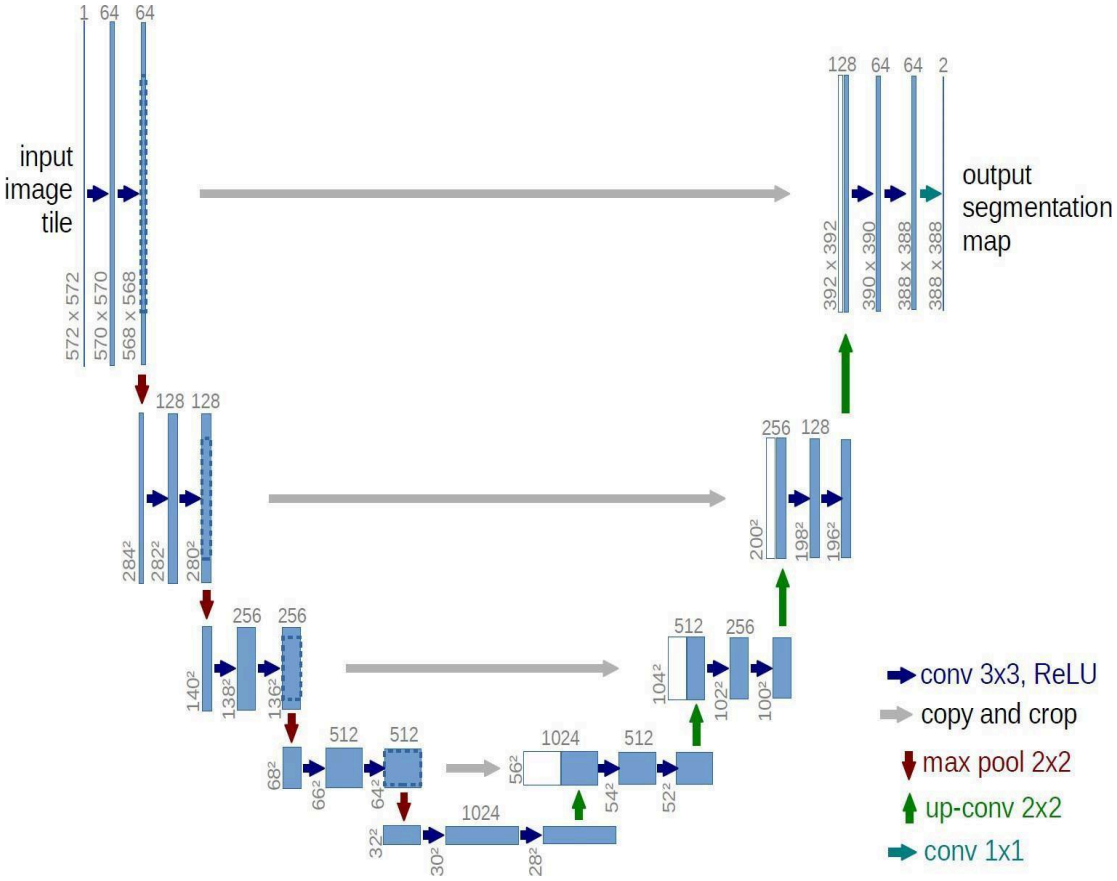


***Figure 2**. The U-Net architecture, an encoder–decoder network designed for biomedical image segmentation, in which the contracting path captures contextual information and the expanding path enables precise localization through skip connections (Ronneberger et al., 2015).*

During the process of pre-training, model optimization typically reduces a classification loss function such as cross-entropy loss through gradient descent-based learning (Goodfellow et al., 2016; LeCun et al., 2015). The learned weights steadily encode hierarchical feature representations, beginning with low-level spatial filters in earlier convolutional layers and slowly moving towards higher level semantic abstractions in the deeper layers (Zeiler & Fergus, 2014; LeCun et al., 2015). For classification tasks, optimization is commonly performed using the categorical cross entropy loss functions, which measures the differences between the predicted probabilities and the ground truth levels.

In medical imaging, the transferability of these learned representations depends strongly upon the resemblance between the source and target domains (Tajbakhsh et al., 2016; Raghu et al., 2019). When pre-trained on natural image datasets, optimization trajectories may become biased toward non-clinical visual features, limiting the ability of the model to identify pathology-specific structures (Raghu et al., 2019; Tajbakhsh et al., 2016). This representation mismatch becomes especially important in volumetric imaging modalities such as CT and MRI, where the contextual continuity across slices and subtle tissue-intensity relationships are clinically meaningful. Domain-specific pre-training somewhat addresses this issue by disclosing the optimization process to the medically relevant spatial and anatomical patterns during representation learning (Azizi et al., 2021; Raghu et al., 2019).

Consequently, the latent feature space learned by the model becomes more parallel with the clinically meaningful variations instead of generic object-recognition features. On the other hand, the domain-specific pre-training enables models to learn the feature representations which are directly aligned with medical imaging features. For example, 3D CNNs which are trained on volumetric CT data can properly capture inter-slice spatial continuity that is clinically very important for anatomical interpretation (Çiçek et al., 2016). Similarly, MRI-focused models can learn relationships across multiple imaging sequences, thereby improving tissue characterization (Isensee et al., 2021). Ultrasound-specific models have also been shown to improve robustness to speckle noises and operator-dependent variability when trained on domain relevant datasets (Liu et al., 2019).

The effectiveness of architectures depend strongly upon the alignment between the training data and the target clinical task. The case where the pre-training data closely matches the modality and task distribution, the models tend to perform well. This allows them to learn meaningful representations. Conversely, when there is a mismatch in the domain, the performance degrades which leads to poor generalization, increased false positives and unreliable clinical predictions.

Therefore, it is the combination of architecture design and domain-specific pre-training that largely determines the success of AI systems in the field of medical imaging rather than the architecture alone (Raghu et al., 2019; Tajbakhsh et al., 2016; Azizi et al., 2021).

The advantages of domain-specific pre-training are also influenced by the dataset size and task complexity. Raghu et al. (2019) showed that ImageNet pre-training may provide modest improvements when medical datasets are small, but the advantage decreases gradually as task-specific medical datasets become larger. This indicates that the transfer learning effectiveness not only depends on architecture selection, but also on the dataset scale, modality complexity, and feature similarity between source and target domains.

Methodology

Data sources - This study is a narrative review that synthesizes existing literature on the role of domain-specific pre-training of AI models in medical imaging. Relevant studies were identified through searches conducted on databases including PubMed, IEEE Xplore, and arXiv.

Search strategy - Keywords were used including, “medical imaging”, “transfer learning”, “domain-specific pre-training”, “MRI deep learning”, “CT segmentation”, “ultrasound AI”, and “chest x-ray deep learning”. Widely cited papers were identified through database citation metrics and the examination of reference lists from highly relevant review articles and foundational studies in medical imaging AI. Full-text versions of potentially relevant studies were manually reviewed to assess the methodological quality, empirical validation, and relevance to the objectives of this review.

Inclusion and exclusion criteria – Peer-reviewed studies and widely cited research papers that provided empirical evaluation of AI models in medical imaging were prioritized. Included studies directly compared domain-specific pre-training with general-purpose pre-training under comparable experimental settings, including image classification, segmentation, and detection tasks. Widely cited papers were selected when they represented foundational contributions to the field or were frequently referenced in subsequent medical imaging literature. Studies that mainly discussed theoretical concepts without empirical validation or did not clearly describe the datasets, evaluation metrics, or model training procedures were excluded from detailed analysis. Studies which compared domain-specific pre-training with general-purpose pre-training were included and discussed on the basis of modality-specific challenges throughout this paper. Methodological clarity was assessed based on the availability of information regarding datasets, model architectures, evaluation protocols, and the reported performance metrics. Studies published between 2015 and 2023 were selected to reflect recent developments in DL.

Study scope and analysis - This review is qualitative in nature and does not perform a formal meta-analysis. Instead, performance metrics reported in the original studies were interpreted within the context of their datasets and experimental conditions. Area Under the Receiver Operating Characteristic Curve (AUROC) was used to assess the discriminative performance in classification tasks because it measures an AI model's ability to distinguish between positive and negative cases across different decision thresholds. However, AUROC alone may not be sufficient for direct cross-study comparisons because datasets, disease prevalence, class distributions, and evaluation protocols can differ substantially between studies. The Dice Similarity Coefficient (DSC) was used to evaluate overlap accuracy in segmentation tasks. Additional metrics such as sensitivity and specificity were considered where reported. Performance comparisons across studies were interpreted cautiously because differences in datasets, class distributions, disease prevalence, evaluation protocols, and experimental settings can substantially influence reported performance metrics, including AUROC and DSC. MRI sequence terminology is also discussed throughout this review, where T1-weighted images primarily provide anatomical detail, T2-weighted images highlight fluid-containing structures and pathology and the contrast-enhanced T1-weighted (T1ce) images improve visualization of lesions and tumor boundaries (Bitar et al., 2006).

Results

The findings across the reviewed studies generally indicated that domain-specific pre-training was associated with improved performance, robustness, and clinical applicability of the AI models across different imaging modalities. Table 1 summarizes the key imaging modalities, associated AI tasks, and their corresponding challenges. Tables 2–5 further compare representative studies across CXR, CT, MRI, and ultrasound imaging modalities, highlighting the differences in robustness, sensitivity, and generalizability between domain-specific and general-purpose pre-trained AI models.

Table 1: Overview of various imaging modalities

Modality	Key AI tasks	Example domain-specific AI models	Challenges addressed
CXR	Pneumonia detection, multi-disease classification	CheXNet	Overlapping anatomical structures, subtle radiographic findings
CT	Lesion detection, volumetric segmentation	3D CNN, 3D U-Net	Volumetric context, scanner variability
MRI	Tumor segmentation, tumor classification	nnU-Net (self-configuring U-Net framework)	Multi-sequence data, tissue-contrast variability
Ultrasound	Organ segmentation, echocardiography analysis	Ultrasound-specific CNNs	Speckle noise, operator dependency

Abbreviations: AI, artificial intelligence; CXR, chest x-ray; CT, computed tomography; 3D, 3-dimensional; CNN, convolutional neural network; MRI, magnetic resonance imaging; nnU-Net, no-new-Net.

Hardware workings of the various medical imaging modalities and their limitations

The CXR imaging is primarily based on the photons passing through the tissues of various densities present (Bushberg et al., 2012; Herring, 2020). While this method is fast and cost-effective, there are some limitations, which include the X-rays having many 2D projections with multiple overlapping anatomical structures, which limit the depth perception and also the sensitivity factors to subtle pathologies such as early COVID-19 infiltrates and other kinds of severe lung diseases (Jacobi et al., 2020; Herring, 2020).

CT is widely used currently in the healthcare industry. The CT scans reconstruct multiple X-ray projections into volumetric 3-dimensional (3D) representations using various advanced tomographic reconstruction algorithms and therefore it offers superior resolution and also shows a proper and a better density variation (Bushberg et al., 2012; Kalender, 2011), but it also has a major limitation that is it exposes the patient to a higher amount of radiation doses which is extremely harmful (Brenner & Hall, 2007)**.**

MRI relies completely on the nuclear magnetic resonance principles to capture the tissue-specific contrast through variations in the relaxation timings (T1, T2) and the pulse sequences (Brown & Semelka, 2010; Westbrook et al., 2018). MRI also provides an exceptionally soft tissue contrast but on the other hand it suffers from extremely high cost, long scan times and also scanner-dependent variability (Brown & Semelka, 2010; Westbrook et al., 2018).

Ultrasound imaging, uses high frequency sound waves to generate real-time images (Hoskins et al., 2010). The advantage of this is that it is safe, portable and cheaper compared to the other imaging modalities but it is highly sensitive to speckle noises and operator dependency (Hoskins et al., 2010; Noble & Boukerroui, 2006).

These modality-specific hardware characteristics directly influence the feature distributions available to the AI models, thereby affecting the effectiveness of different pre-training strategies and contributing to variations in model generalizability across the imaging modalities.

Chest X-Ray Imaging

One of the most widely used diagnostic modalities commonly used in clinical practices is the CXR. This is generally used for the detection and also monitoring of pulmonary and cardiovascular diseases. Since the cost is low, the outcome is prompt, and it is also broadly available, chest radiology is considered to be the primary and the first imaging tool for detecting conditions such as pneumonia, tuberculosis, lung cancer, COVID-19, pleural effusion, heart failure.

CXRs present multiple analytical challenges for both automated systems and AI models because the images are two-dimensional (2D), with overlapping anatomical structures and subtle intensity variations that often require expert interpretation. Consequently, CXR analysis remains a complex task, and several studies suggest that domain-specific pre-training enables AI models to learn radiographic features that are more relevant for clinical interpretation than representations learned from natural-image datasets (Raghu et al., 2019; Litjens et al., 2017).

One of the most prominent examples of such training in the case of CXR imaging is *CheXNet,* proposed by Rajpurkar et. al. (2017). Beyond pneumonia, various recent studies and literature have demonstrated the effectiveness of the usage of domain-specific AI models in detecting COVID-19 related things, tuberculosis cavitation and lung nodules as well, which are often subtle and easily missed by the general-purpose AI models. Since it was trained on large-scale expertly labelled X-ray datasets, CheXNet achieved state-of-the-art performance on multiple thoracic diseases, including pneumonia, with AUROC scores typically in the low-to-mid 0.80s on the ChestX-ray14 test set, comparable to radiologist-level performance (Rajpurkar et al., 2017).

It was also found that CheXNet illustrated improved diagnostic performances in several reviewed parameters as compared to the general CNNs. On the other hand, the performances of general-purpose CNNs pre-trained on natural images were stated to show limited generalizability in certain clinical settings (Zech et al., 2018). Zech et al. (2018) also showed that the AI models that were trained on CXRs from a single institution were unable to generalize to external datasets. Especially when pretrained on non-medical images.

These findings suggest that the performance gaps between domain-specific and general-purpose pre-training are strongly associated with feature transferability and the dataset biases. These differences are likely related to the limited transferability of features learned from natural-image datasets to chest radiography, particularly when disease manifestations are subtle (Zech et al., 2018; Roberts et al., 2021). In chest radiography, models sometimes rely on false correlations such as hospital specific markers, scanner artefacts, or even acquisition patterns instead of the pathology-specific anatomical features (Zech et al., 2018; Roberts et al., 2021). This issue becomes very important especially during external validation, where the models trained under narrow institutional distributions may experience substantial drops in performances.

Collectively, these findings suggest that domain-specific pre-training may contribute to an improved diagnostic performance, robustness, and generalizability in CXR analysis when compared to general-purpose pre-training methods. However, the magnitude of these benefits may vary greatly depending on the dataset characteristics, model architecture, and evaluation settings. There are more recent studies that comparatively evaluated and found that the models pretrained on large-scale CXR datasets most of the times performed much better when compared to the general-purpose vision models of comparable sizes in both the classification and segmentation tasks (Irvin et al., 2019).

Domain-specific pre-training may also improve model interpretability by encouraging AI systems to focus on clinically relevant anatomical regions rather than non-diagnostic image artefacts. Several explainability studies using gradient-based localization techniques have shown that domain-trained AI models more consistently highlight pulmonary opacities, nodules, and other diagnostically relevant thoracic regions (Rajpurkar et al., 2017; Roberts et al., 2021).

These findings indicate that both the foundational and the more recent studies consistently report potential benefits of using domain-specific pre-training for CXR analysis. Taken together, the evidence which is available suggests that the modality aware pre-training may play a crucial role in improving the clinical reliability of AI systems used in chest radiography. The importance of domain-specific pre-training is further supported by the availability of large-scale medical imaging datasets. For example, the CheXpert dataset contains 224,316 chest radiographs from 65,240 patients, providing substantial modality-specific data for representation learning and model development (Irvin et al., 2019). The DL models which were trained on CXR images specifically achieved a higher multi-pathology classification performance when compared to the general foundation AI models.

Table 2 summarizes representative comparisons between CXR specific models and ImageNet-pretrained CNNs across pneumonia classification and multi-disease detection tasks.

Table 2: Chest X-Ray AI Model Comparison

Model	Pre-training	Task	Performance	Notes
CheXNet	CXR dataset	Pneumonia classification	AUROC ≈ 0.82–0.85	Evaluated on ChestX-ray14 dataset (Rajpurkar et al., 2017)
CNN	ImageNet	Pneumonia classification	Performance varies across datasets	Reduced external generalizability reported (Zech et al., 2018)
CheXNet	CXR dataset	Multi-disease detection	Strong performance across multiple thoracic diseases	Supported by multiple studies

Domain-specific models such as CheXNet demonstrate improved robustness and generalizability compared to general-purpose CNNs, particularly under cross-institutional evaluation settings. Abbreviations: CXR, chest x-ray; AUROC, Area Under the Receiver Operating Characteristic Curve; CNN, convolutional neural network.

Computed Tomography Imaging

CT imaging plays a vital role in day to day clinical workflows, especially in oncology and trauma assessment tests (Kalender, 2011; Bushberg et al., 2012). CT scan imaging is also widely used in the detection of pulmonary embolism, strokes, intracranial hemorrhages, and COVID-19 related abnormalities (Rubin et al., 2015; Kwee & Kwee, 2020). CT scans provide a high-resolution 3D volumetric data which properly capture finely grained anatomical structures and details across a wide range of slices (Kalender, 2011). These characteristics demand AI models that are capable of understanding complex 3D spatial relationships, tissue-density variations, and subtle pathological patterns (Çiçek et al., 2016; Litjens et al., 2017).

CT imaging represents one of the most complex modalities for reliable automated medical analysis. General purpose vision models trained on two dimensional (2D) natural images often struggle to interpret volumetric CT data effectively because they are designed primarily for planar image representations rather than volumetric feature learning (Raghu et al., 2019; Tajbakhsh et al., 2016). These models therefore lack intrinsic mechanisms for learning spatial continuity across adjacent slices and may thereby fail to properly capture clinically relevant 3D contextual information (Çiçek et al., 2016; Isensee et al., 2021).

In contrast, the AI models trained on volumetric CT data demonstrate highly improved sensitivity and robustness in the clinical detection and segmentation tasks (Çiçek et al., 2016; Setio et al., 2016). Previous studies demonstrate that CT specific models which are trained on volumetric data improve cancer detection sensitivity and lesion localization to a great extent when compared to 2D vision models pretrained on natural images (Çiçek et al., 2016; Setio et al., 2016). Several studies show high improvements in sensitivity for lesion detection when CT-specific 3D architectures are used in comparison to 2D vision models (Setio et al., 2016).

CT-specific domain-trained models have been demonstrated to reduce false positives in lesion detection tasks, which is an extremely important and crucial factor in minimizing unnecessary follow-up procedures (Setio et al., 2016). These findings show the disadvantages of the general-purpose foundational AI models in handling volumetric medical data and also highlight the necessity of modality-aware pre-training.

A major difference between CT-specific architectures and conventional 2D CNN transfer learning approaches lies in their representations of spatial context. Standard ImageNet-pretrained CNNs process slices independently and therefore may lose inter-slice anatomical continuity, on the other hand, 3D CNNs and volumetric U-Net variants learn spatial dependencies directly across adjacent slices. These architectural differences become particularly important for detecting diffuse lesions, irregular tumor boundaries and low-contrast abnormalities that require volumetric contextual interpretation (Çiçek et al., 2016).

Another major advantage of CT-specific domain pre-training is in its ability to account for scanner-related variability and reconstruction artefacts. The CT images vary a lot, depending a lot on the scanner manufacturers, the radiation dose protocols and the reconstruction algorithms. The pre-training of AI models with specific domain data enables the AI models to learn invariant features properly and in depth so that they can generalize across such variations, whereas general-purpose vision models often wrongly interpret reconstruction artefacts as pathological features and hence limit their clinical reliability (Litjens et al., 2017; Roberts et al., 2021).

Domain-specific CT models are not completely without limitations. Although volumetric pre-training improves contextual understanding to a great extent, these models are computationally expensive and require mostly larger annotated datasets compared to the conventional 2D transfer learning approaches. Moreover, cross-institutional variability in CT acquisition protocols may still reduce the external generalizability despite domain-specific training. This suggests that architecture selection itself is not sufficient and must be combined with dataset diversity and thorough external validation.

Overall, the reviewed evidence highlights the limitations of general-purpose vision AI models for CT imaging and supports the use of domain-specific pre-training to improve performance, robustness, and clinical applicability. For complex applications such as cancer detection and trauma diagnosis, domain-specific CT AI models are more appropriate for achieving reliable clinical performance.

Table 3 summarizes representative comparisons between CT-specific 3D architectures and conventional ImageNet-pretrained 2D CNN approaches in volumetric lesion detection tasks.

Table 3: Computed Tomography AI Performance

Model	Pre-training Type	Task	Performance	Notes
3D CNN	CT-specific dataset	Lesion detection	Higher sensitivity reported in controlled studies	Captures volumetric context (Setio et al. (2017))
2D CNN	ImageNet	Lesion detection	Lower sensitivity in volumetric tasks	Limited 3D understanding (Çiçek et al. (2016); Setio et al. (2017))

CT specific AI models such as CNNs achieve a higher performance in volumetric lesion detection tasks because they preserve volumetric special relationships. Abbreviations: 3D, 3-dimensional; CNN, convolutional neural network; CT, computed tomography; 2D, 2-dimensional.

Magnetic Resonance Imaging

MRI has emerged as one of the most widely used imaging tools in modern day medical diagnosis, especially in the fields of neurological imaging, oncology, and musculoskeletal imaging. Unlike other medical imaging tools such as X-ray technology-based medical imaging modalities, MRI image has very high-resolution volumetric data with very subtle contrast in soft tissues that are demonstrated through appropriate image acquisition protocols such as the T1-weighted image, the T2-weighted image, and the contrast-acquired image (T1ce image). This very multi-sequence and high-dimensional nature of the MRI data makes it in particular less acceptable for direct transfer learning from the natural image datasets. The analysis of an MRI image becomes more complex for a general AI system, this is because there is a large variation in the intensities specified to MRI image capturing tools which is again combined with a large amount of anatomical details which were absent in natural images. Due to this, MRI serves as a critical platform for evaluation of the effectiveness of domain-specific pre-training in medical imaging models. The transfer learning from the natural image datasets most of the times provide a slight benefit in capturing the modality-specific intensity variations which are inherent to MRI data (Raghu et al., 2019).

A growing body of the existing literature continuously shows that the pre-trained AI models, those that are pre-trained on specific MRI datasets are much better than the general-purpose foundational AI models of same sizes across a wide range of tasks including tumor detection, classification and segmentation. Raghu et al. (2019) investigated the

effectiveness of features learnt from the datasets which contained natural images in medical imaging tasks and demonstrated that the ImageNet pre-training provided very limited benefits for applications which were based on MRI. Their findings showed that when the CNN architectures of the similar sizes and capacities were compared, AI models pre-trained directly on MRI datasets consistently outperformed ImageNet-pretrained counterparts across medical imaging tasks such as brain tumor detection and tumor segmentation (Raghu et al., 2019). This limitation arises because MRI data contains modality-specific features, which includes tissue contrast mechanisms, multi-sequence information, and complex anatomical structures which are not present in the natural-image datasets. Consequently, representations learned from natural images may not adequately capture clinically relevant MRI characteristics, thereby limiting the effectiveness of direct transfer learning without domain-specific exposure (Raghu et al., 2019; Tajbakhsh et al., 2016; Litjens et al., 2017). These findings regarding the limited transferability of the natural-image pre-training are further supported by studies which demonstrate that the AI models which are trained directly on MRI data achieve superior tumor segmentation performances across multiple MRI sequences, including T1-weighted, T2-weighted, and contrast-enhanced T1-weighted (T1ce) images, when compared to the AI models initialized using natural-image datasets (Bakas et al., 2018; Isensee et al., 2021). This advantage in the performance has been continuously observed on public brain tumor segmentation benchmarks such as the Brain Tumor Segmentation (BraTS) Challenge (Bakas et al., 2018). Across BraTS benchmark evaluations, MRI-specific training strategies such as nnU-Net have consistently reported Dice similarity coefficients exceeding 0.80 for several tumor sub-regions, depending on the segmentation task and validation protocol (Isensee et al., 2021). On the contrary, evidence from transfer-learning studies suggests that ImageNet pre-training provides only a limited number of benefits for MRI-based segmentation tasks because the learned representations do not properly capture the MRI-specific tissue characteristics and anatomical structures (Raghu et al., 2019). These findings therefore highlight the limitations of relying solely on natural-image pre-training for MRI segmentation applications.

Cross-study comparisons show that the benefits of MRI-specific pre-training can become more pronounced as task complexity increases. While ImageNet transfer learning may provide moderate initialization advantages for simpler classification tasks with limited datasets, segmentation tasks which involve heterogeneous tumor boundaries and multi-sequence fusion will generally require modality-specific representational learning. This suggests that the effectiveness of transfer learning depends greatly upon the interaction between the task complexity, dataset scale, and modality-specific structural variation (Raghu et al., 2019; Isensee et al., 2021). These differences in performance are particularly significant in clinical matters, where proper and precise tumor boundary demarcation directly influences the treatment planning and the diagnostic outcomes. In addition to these segmentation tasks, the MRI-specific foundational AI models have also demonstrated a much better performance in classification tasks like tumor grading and the detection of diseases. These tasks include stroke outcome prediction as well.

Multiple studies and literature have reported that the MRI-specific representational learning improves the disease classification and also the tumor grading performances when compared to the general-purpose foundational vision AI models, especially when the model sizes are the same (Raghu et al., 2019; Litjens et al., 2017). These improvements show the importance of AI models learning MRI-specific anatomical priors, which the general vision AI models often fail to capture effectively and properly. Other important and crucial factors influencing the MRI performances are scanner variability and acquisition heterogeneity. MRI images vary greatly across the scanner manufacturers, field strengths and patient demographics. Domain-specific pre-training with proper and wide MRI datasets exposes the AI models to the scanner and acquisition variability during learning, which leads to improved robustness and cross-institutional generalization, in contrast, the general-purpose models often fail to adapt reliably and effectively to the unseen MRI distributions (Raghu et al., 2019; Roberts et al., 2021). Such robustness is extremely crucial for the safety and the reliability factor of clinical deployment of the MRI based AI models.

Some studies also indicate that domain-specific pre-training itself does not completely eliminate generalization failures. The differences in MRI acquisition protocols, magnetic field strengths, and institutional imaging practices may still produce significant performance variability during external validation. Therefore, although MRI-specific pre-training improves representation quality, robust clinical deployment additionally requires heterogeneous datasets, standardized evaluation protocols and rigorous external benchmarking as well (Roberts et al., 2021).

Overall, the extensive empirical evidence confirms that the MRI-based tasks substantially benefit from domain-specific pre-training with MRI-specific data when the AI models of the similar sizes are compared. These findings properly highlight that the MRI's modality-specific complexity requires proper and appropriate learning strategies and that the reliable clinical deployment of AI in MRI cannot rely solely on the general-purpose vision pre-training.

Table 4 summarizes representative comparisons between MRI-specific architectures and ImageNet-pretrained CNN approaches across segmentation tasks on BRATS and related benchmarks.

Table 4: Magnetic Resonance Imaging AI Performance

Model	Pre-training Type	Task	DSC / AUROC	Notes
nnU-Net	MRI dataset	Brain tumor segmentation	DSC ≈ 0.80+ to 0.90+	BRaTS benchmark (Isensee et al., 2021)
CNN	ImageNet	Brain tumor segmentation	Typically, lower than MRI specific models	Limited transferability (Raghu et al. (2019))

MRI specific AI models have frequently been reported to achieve a higher performance than ImageNet pretrained AI models because they learn modality specific tissue contrast and multi-sequence dependencies. Abbreviations: DSC, Dice Similarity Coefficient; AUROC, Area Under the Receiver Operating Characteristic Curve; nnU-Net, no-new-Net; MRI, Magnetic Resonance Imaging; BRaTS, Brain Tumor Segmentation Challenge; CNN, convolutional neural network.

Ultrasound Imaging

Ultrasound imaging, due to its non-invasiveness, real-time acquisition and low cost, is being widely adopted across a wide range of clinical domains (e.g., abdominal imaging, obstetrics and cardiology). Other important applications include vascular assessment and echocardiography assessments as well. However, ultrasound imaging carries a number of complex and unique challenges, e.g., low signal-to-noise ratio, speckle noises, operator dependency, and considerable variability in image qualities. The general foundational vision AI models which are trained on natural image datasets may misinterpret speckle noises and acquisition artefacts as meaningful features, thereby leading to unreliable predictions (Xie et al., 2018; Chen et al., 2019). On the other hand, domain-specific pre-training on large-scale ultrasound data enables AI models to learn robust clinically relevant patterns despite the substantial noise and operator variability (Xie et al., 2018; Chen et al., 2019). Multiple surveys in the literature have shown that DL models pre-trained on domain-specific ultrasound data consistently yield strong performance compared to general foundational AI models with comparable parameter sizes (Xie et al., 2018; Chen et al., 2019).

Performance metrics which include accuracy (the proportion of correctly classified cases), F1 score (the harmonic mean of precision and recall) and robustness under noisy imaging conditions generally favor the domain-trained ultrasound AI models. To be specific, under noisy and operator-dependent imaging conditions, AI models pre-trained on domain-specific ultrasound data most of the times outperform general-purpose foundational AI models of comparable size (Xie et al., 2018; Chen et al., 2019).

Unlike CT and MRI, ultrasound imaging introduces a significant operator-dependent variability during image acquisition. Consequently, performance differences between domain-specific pre-training and general-purpose pre-training are not only influenced by representation learning, but also by acquisition inconsistency. Domain-specific ultrasound models are generally considered better at learning invariant features under noisy imaging conditions, while natural-image pretrained models may incorrectly amplify speckle artefacts or acquisition distortions during feature extraction. These differences are very critical when AI is taken and applied under real-world clinical settings, especially when ultrasound images vary significantly across devices and patient populations. This significant performance gap between general-purpose foundational vision AI models and ultrasound-specific AI models accentuates the need for modality-aware training in AI models in the healthcare field for trustworthy, accurate, and reliable results.

The literature surrounding ultrasound AI remains comparatively less standardized than MRI and CT research. Dataset sizes are often smaller, annotation quality varies substantially, and benchmarking protocols are less uniform across the studies. As a result, although current evidence strongly supports the advantages of ultrasound-specific pre-training, the direct comparison across studies remain difficult because of the differences in acquisition settings, devices, and operator

expertise.

Table 5 summarizes representative comparisons between ultrasound-specific CNNs and ImageNet-pretrained CNN models across noisy and operator-variable clinical imaging conditions.

Table 5: Ultrasound AI Performance

Model	Pre-training Type	Task	Performance	Notes
Ultrasound-specific CNN	Ultrasound dataset	Clinical tasks	Improved robustness under noise conditions	Domain adaptation beneficial (Xie et al. (2018); Chen et al. (2019))
CNN	ImageNet	Clinical tasks	Performance sensitive to noise and variability	Less robust (Xie et al. (2018); Chen et al. (2019))

Studies reviewed in this paper suggest that domain-specific ultrasound AI models provide a greater robustness to noise and operator variability than general-purpose AI models. Abbreviations: CNN, convolutional neural network.

Discussion

Across the various imaging modalities discussed in this review, the literature generally suggests that the AI models which are pre-trained on domain-specific medical datasets are associated with improved performance, robustness, and clinical applicability when compared to general-purpose foundational AI models (Tajbakhsh et al., 2016; Raghu et al., 2019; Zech et al., 2018; Roberts et al., 2021). Various findings from medical imaging studies suggest that modality specific details and pathologies are often absent in the natural image datasets, which may contribute to reduced performance in certain clinical tasks. Collectively, the reviewed studies indicate that the alignment between the pre-training dataset and the target imaging modality plays a vital role in determining feature transferability and downstream clinical performance (Tajbakhsh et al., 2016; Raghu et al., 2019; Zech et al., 2018; Oakden-Rayner, 2020; Roberts et al., 2021). Sometimes, general vision AI models may also have difficulty adapting to variations in image acquisition conditions, image reconstruction algorithms, and patient demographics as well. These variances are more effectively accommodated by domain-specific pre-training, which may contribute to more reliable clinical predictions.

The reviewed literature also indicates that the magnitude of improvement associated with domain-specific pre-training is not uniform across all imaging modalities. Simpler classification tasks with limited datasets might sometimes still benefit moderately from ImageNet initialization, especially when computational resources or annotated medical data are limited (Tajbakhsh et al., 2016). However, as task complexity increases, specifically in volumetric segmentation, multi-sequence analysis, or noisy imaging environments, the limitations of natural-image transfer learning becomes more apparent. Across the studies included in this review, CT and MRI applications generally reported larger improvements than CXR classification tasks because volumetric imaging and multi-sequence analysis require richer modality-specific feature representations. Similarly, ultrasound imaging benefited from domain-specific pre-training through improved robustness to speckle noise, operator dependency and acquisition variability. These findings suggest that the advantages of domain-specific pre-training becomes more pronounced as imaging complexity and contextual requirements increase.

The domain-specific image datasets have also shown considerable scope for further improvement. There are widely used image datasets that are not fully expert-verified by radiologists, which may result in residual annotation variability (Rajpurkar et al., 2017). In addition, some image datasets also lack adequate demographic diversity, which may raise concerns regarding fairness and model generalizability (Oakden-Rayner, 2020; Roberts et al., 2021). To overcome these constraints, continued collaboration between clinicians, radiologists, and AI researchers is essential. Domain-specific AI models are not intended to replace doctors or radiologists. Instead, they serve as decision-support tools that can enhance human-AI collaboration. When these AI models are developed using appropriately curated datasets and rigorous validation, they may function as a reliable second reader, helping to reduce radiologists' workload and support clinical decision-making. Nevertheless, the reviewed literature suggests that robust external validation and diverse,

expert-annotated datasets remain essential for achieving clinically reliable AI systems. Though there are substantial advantages of domain-specific pre-training, there are certain limitations as well.

One major challenge is the limited availability of large-scale, high-quality annotated medical datasets, because expert annotation is both time-consuming and also very expensive. Moreover, many existing datasets lacks sufficient demographic diversity, which may lead to bias and limit the generalizability of AI models across different populations (Roberts et al., 2021; Oakden-Rayner, 2020). Furthermore, variability in imaging protocols, scanner types, and acquisition settings across institutions can affect model robustness despite domain-specific pre-training. The computational costs related to training large-scale domain-specific models also remain high, thereby posing practical constraints for widespread adoption.

When compared with alternative approaches such as transfer learning and emerging techniques such as meta-learning, domain-specific pre-training often provides better alignment between learned representations and modality-specific clinical features. Transfer learning enables adaptation of previously learned representations to new tasks using limited labelled data, whereas meta-learning aims to improve rapid generalization across multiple tasks and domains (Pan & Yang, 2010; Hospedales et al., 2021). However, several studies suggest that modality-specific pre-training may capture clinically better relevant imaging characteristics when sufficient domain data are available (Raghu et al., 2019; Tajbakhsh et al., 2016). Future research should therefore explore hybrid strategies that combine domain-specific representation learning with adaptable learning frameworks to improve both scalability and generalization (Hospedales et al., 2021).

Recent literature suggests that self-supervised learning may become an important complementary strategy in medical imaging. Unlike conventional supervised pre-training, self-supervised approaches learn representations directly from large volumes of unlabelled medical data through pretext or proxy tasks, thereby reducing dependence on expensive expert annotations (Azizi et al., 2021; Taleb et al., 2020). Several studies have reported that self-supervised domain-specific pre-training can improve robustness, feature transferability, and downstream task performance, particularly in data-limited clinical environments (Azizi et al., 2021; Taleb et al., 2020). However, external validation procedures and benchmarking frameworks remain insufficiently standardized across institutions, and further comparative research is required before these approaches can be integrated reliably into routine clinical workflows (Roberts et al., 2021).

Overall, the findings synthesized in this review suggests that the effectiveness of pre-training strategies depends on the interaction between imaging modality, dataset characteristics, task complexity and evaluation methodology. Rather than relying on a single universal pre-training strategy, future medical imaging systems are likely to benefit from modality-aware learning frameworks supported by diverse datasets, rigorous external validation, and clinically meaningful evaluation protocols.

Future Directions

Future research in medical imaging should focus not only on improving predictive performance but also on enhancing clinical reliability, external validation, and explainability across multiple healthcare settings. One important direction involves the development of large-scale multi-institutional datasets that encompass demographic diversity, heterogeneous scanner protocols, and geographically distributed patient populations. Such datasets would help reduce dataset-specific bias while improving model robustness during external validation (Roberts et al., 2021).

Another important research direction involves the development of multimodal learning frameworks that combine information from multiple imaging modalities, electronic health records, pathology reports and other relevant clinical metadata as well. Existing literature suggests that multimodal integration may improve contextual reasoning and reduce reliance on isolated image features, thereby potentially enhancing diagnostic consistency in complex clinical settings (Litjens et al., 2017).

Future work should also prioritize explainability and uncertainty estimation in clinical AI systems. Although high AUROC or DSC values may indicate strong benchmark performance, these metrics alone do not guarantee clinically reliable decision-making. Explainable AI techniques, saliency mapping methods, and uncertainty aware prediction frameworks may assist clinicians in interpreting model outputs more effectively while identifying predictions that require additional clinical review.

Greater systematic comparisons are also needed between domain-specific pre-training, transfer learning, meta-learning, and self-supervised learning under comparable evaluation conditions. Current literature often evaluates these approaches on different datasets and experimental settings, making direct comparison more difficult. Benchmarking frameworks that incorporate common datasets, external validation protocols, and clinically meaningful evaluation metrics would thereby improve reproducibility facilitate more reliable assessment of these learning strategies.

Finally, future clinical deployment research should prioritize prospective validation in real-world clinical environments rather than relying primarily on retrospective benchmark datasets. Several studies have shown that AI systems performing strongly in controlled experimental settings may sometimes fail under real clinical variability because of distribution shifts, demographic imbalance, or institutional differences (Roberts et al., 2021). Consequently, clinically deployable AI systems are likely to require continuous monitoring, recalibration, and collaborative oversight involving clinicians, radiologists and AI researchers to ensure sustained safety, reliability, and clinical effectiveness.

Conclusion

Domain-specific pre-training plays a vital role in improving the performance, robustness, and clinical applicability of AI models across various medical imaging modalities. The effectiveness of pre-training strategies depends on factors including modality characteristics, dataset scale, task complexity and external validation conditions as well. When compared to general-purpose vision AI models pre-trained on natural-image datasets, the domain-specific pre-training often enables the learning of representations that are more closely aligned with clinically relevant anatomical and pathological features.

The findings presented for CXR, CT, MRI and ultrasound applications highlight the potential advantages of modality-aware learning strategies for supporting proper and reliable medical image analysis. However, clinically reliable deployment requires more than just improved model performance alone. Diverse and representative datasets, rigorous external validation, standardized benchmarking frameworks, explainable AI methods and prospective real-world evaluation remain essential for improving safety and generalizability.

Future research should focus on multimodal data integration, cross-institutional validation, improved dataset diversity, systematic comparisons between domain-specific pre-training and alternative learning paradigms, and the development of clinically deployable AI systems that support effective human-AI collaboration as well. Taken together, these efforts may contribute to the development of more reliable, transparent and clinically useful AI systems for medical imaging.

Acknowledgement

The author of this paper acknowledges the contributions of all the researchers whose work formed the foundation of this review paper and also extends his sincere gratitude to his mentors and peers for their valuable guidance and feedback all throughout this research process.

Bibliography

1. 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.
2. Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems, 32.
3. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225.
4. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine, 15(11), e1002683.
5. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019, July). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 590-597).
6. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods, 18(2), 203-211.
7. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., ... & Jambawalikar, S. R. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629.
8. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning?. IEEE transactions on medical imaging, 35(5), 1299-1312.
9. Setio, A. A. A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., Van Riel, S. J., ... & Van Ginneken, B. (2016). Pulmonary

nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE transactions on medical imaging, 35(5), 1160-1169.
10. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016, October). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In International conference on medical image computing and computer-assisted intervention (pp. 424-432). Cham: Springer International Publishing.
11. Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S. X., ... & Wang, T. (2019). Deep learning in medical ultrasound analysis: a review. Engineering, 5(2), 261-275.
12. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC medicine, 17(1), 195.
13. Oakden-Rayner, L. (2020). Exploring large-scale public medical image datasets. Academic radiology, 27(1), 106-112.
14. Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., ... & Schönlieb, C. B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence, 3(3), 199-217.
15. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (pp. 234–241). Cham: Springer.
16. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., ... & Corrado, G. (2021). Big Self-Supervised Models Advance Medical Image Classification. Proceedings of ICCV 2021.
17. Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2021). Meta-learning in neural networks: A survey. IEEE TPAMI.

Enhancing the clinical reliability of AI models in medical imaging through the use of domain-specific pre-training across various medical imaging modalities

Background: Artificial Intelligence (AI) has been integrated into many sectors, leading to which has led to improvements in accuracy and efficiency. AI has demonstrated highly substantial improvements in image classification, object detection, and image segmentation tasks in which they oftentimes can outperform the conventional approaches under controlled conditions. Healthcare is one of the most critical domains because even minor diagnostic errorsthe slightest of errors can significantly affect patient safety and treatment planningin any kind of diagnosis can highly impact the safety of the patient and treatment planning. Therefore, reliability is one of the most important factors for any AI system operating in the healthcare sector. Despite this, many AI models, which are primarily pre-trained on natural images, are used in various healthcare applications, which limits may limit their clinical reliability and generalizability.

Methods: This manuscript presents a qualitative narrative review of existing literature, comparing general-purpose pre-training strategies and domain-specific pre-training strategies across four major imaging modalities: which are, namely, chest X-rays (CXR), magnetic resonance imaging (MRI), computed tomography (CT), and ultrasound. Relevant studies were identified through searches across major databases, including PubMed, IEEE Xplore and arXiv, using inclusion and exclusion criteria which is defined in the methodology section. This review is qualitative in nature and does not perform a formal meta-analysis.

Results: The reviewed studies generally suggest that AI models pre-trained on domain-specific data demonstrate an improved performance, robustness, and clinical applicability across multiple medical imaging modalities when compared with AI models that are pre-trained on natural image datasetsacross all modalities d.The AI models that are trained on domain-specific data consistently demonstrate superior performance and robustness in clinically applicable tasks, as reported across multiple studies in terms of performance, robustness, and clinical applicability when compared to AI models which are pre-trained on natural image datasets.

Conclusion: The reviewed evidence suggests that domain-specific pre-training represents a very promising strategy for improving the clinical reliability and generalizability of AI systems used in in the field of medical imaging. Future research should focus on multimodal multi-modal data integration, cross-institutional validation, real-world clinical testing, and systematic comparison witcomparisons withh alternative learning paradigms such as transfer learning and meta-learning.

Keywords: artificial intelligenceAI, medical imaging, domain-specific pre-training, CXRchest x-ray, MRImagnetic resonance imaging, computed tomographyCT, ultrasound, human-AI collaboration

Introduction

Medical imaging is vital for modern clinical diagnosis, thereby allowing, allowing for non-disruptive imaging of anatomical structures and, any abnormal bodily function of cells/ and tissues, required for treatment, prognosis, and disease mapping. Imaging modalities such as chest x-ray (CXR), computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound are widely used in the healthcare sector and are a critical component of the diagnostic workflow.

In recent years, artificial intelligence (AI), particularly deep learning (DL) (DL), has been a significant tool for medical image analysis, allowing noticeable improvements in diagnostic accuracy, efficiency, and clinical progress (Litjens et al., 2017). The success of DL in general computer vision, which includes tasks such as image classification, object detection, and object recognition using natural-image datasets, has accelerated the adoption of general-purpose foundational AI models in medical imagingThe success of deep learningDL in general computer vision has quickened the adoption of general-purpose foundational AI models in medical imaging. These AI models are, mostly of which are pre-trained on large-scale natural image datasets. Large-scale natural-image datasets include ImageNet, which is the most widely used dataset for pre-training DL models in computer visionLarge-scale natural image datasets are XYZ. These AI models in medical imaging have shown exceptional promising accomplishment across a wide range of general visual tasks, which motivated their shift to medical imaging applications. As medical images are modality-specific, they greatly differ greatly from natural images and thus require trained and expert clinical data interpretation. Consequently, AI models trained on the basis of non-medical data often may not be generalized optimally forto medical imaging tasks, potentially limiting their

clinical reliability and robustness ~~fail to show effective results in the clinical field or practice. This raises concern regarding safety, accuracy, and dependability~~ (Kelly et al., 2019; Roberts et al., 2021). Previous studies suggest that models pretrained on natural-image datasets may not adequately capture clinically relevant anatomical and pathological features, which can reduce performance on certain medical imaging tasks (Tajbakhsh et al., 2016; Litjens et al., 2017).

~~Every~~ All imaging modalit~~ies~~y face~~s~~ challenges and difficulties and thus demands domain-aware learning. ~~In the case of CXR, there are two-dimensional representations with overlapping anatomical structures that require refined layered interpretation of distinct regions of the lungs, cardiac profile, and delicate radiographic patterns for the identification of diseases such as pneumonia and tuberculosis~~. MRI offers superior soft-tissue contrast through multiple imaging sequences, thereby enabling comprehensive assessment of neurological and musculoskeletal abnormalities, but requires an understanding of complex tissue-specific signal intensity variations~~MRI offers a superior soft-tissue contrast through multiple phase sequences (e.g., T1, T2, T1ce), thereby enabling a comprehensive understanding of neurological and musculature abnormalities, but requires~~~~necessitating an understanding of complicated tissue-specific~~. CT provides high-resolution volumetric imaging for applications such as oncology and trauma assessment~~, while u~~. Meanwhile, ultrasound presents ~~additional~~ challenges ~~due to~~such as speckle noise, operator dependency, and variable image quality. ~~G~~However, general foundational AI models are still used for many of these imaging tasks despite these modality-specific challenges. ~~Previous studies suggest that models pretrained on natural-image datasets may not adequately capture clinically relevant anatomical and pathological features, which can reduce performance on certain medical imaging tasks~~ magnitude variations. ~~General foundational AI models are still used in spite of such challenges. With the help of large-scale surveys, it has been found that general foundational AI models are often not accurate and have inadequate key features like clinical bases and lead to wrong explanations of anatomical features (Tajbakhsh et al., 2016; Litjens et al., 2017). A comparison across existing literature demonstrates the advantages of domain-specific pre-training in improving model robustness and clinical relevance (Litjens et al., 2017; Raghu et al., 2019).~~ ¶

The process of training an AI model on a large dataset to learn general feature representations before fine-tuning it on a specific task is referred to as "pre-training". In medical imaging, general-purpose pre-training using natural-image datasets such as ImageNet often ~~on foundational AI models (e.g. ImageNet) mostly~~ provides limited ~~an insubstantial~~ benefits due to the fundamental differences between natural and medical images~~, which include modality- specific intensity distributions and anatomical structures~~ (Raghu et al., 2019).¶

¶

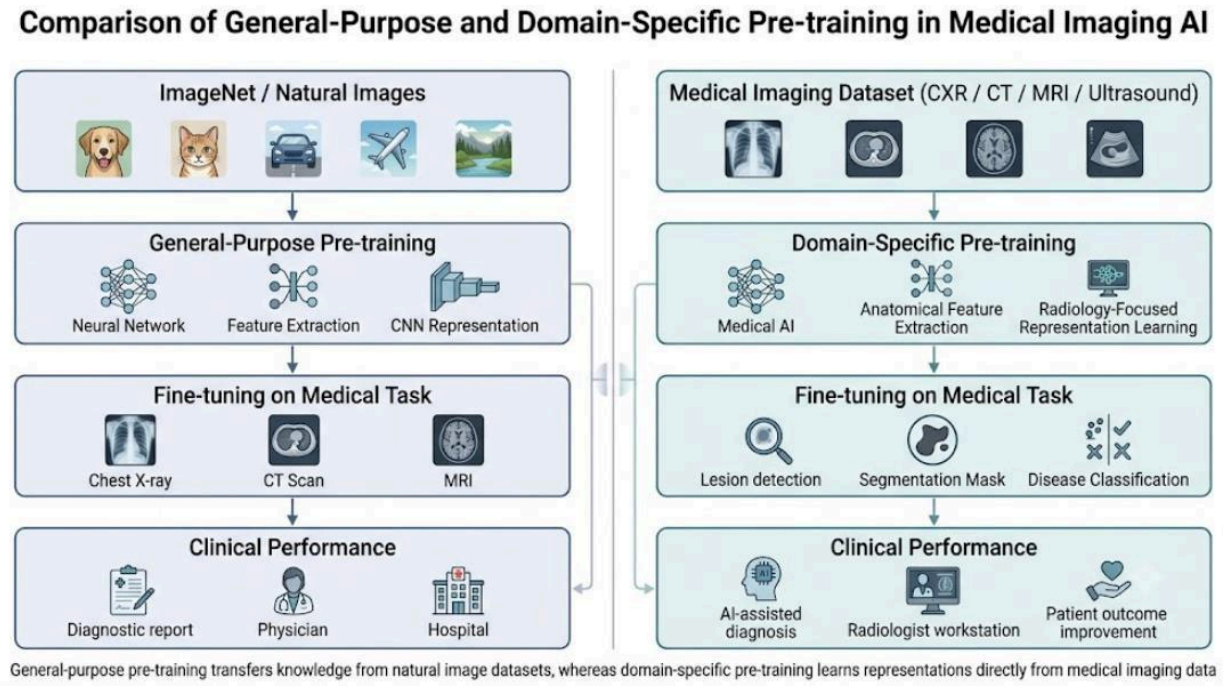
The limited transferability of natural-image pre-training has been associated with ~~to~~ differences in feature distributions and optimization behavior between ~~the~~ medical imaging datasets ~~medical~~ and non-medical datasets. In ~~the~~ natural-image datasets such as ImageNet, convolutional neural networks (CNNs) primarily learn edges, textures, colors, and object-level semantic features. However, medical imaging tasks oftentimes depend upon ~~the~~ subtle grayscale intensity variations, tissue morphology and pathology-specific structures that differ greatly from natural images ~~figures~~ (Raghu et al., 2019; Tajbakhsh et al., 2016). As a result, pretrained representations derived from ~~the~~ natural image datasets ~~s~~ may ~~merge~~ converge towards clinically irrelevant features, thereby resulting in weaker generalization under cross-institutional or modality-shifted conditions. Domain-specific pre-training addresses these limitations by exposing AI models to medically relevant anatomical and pathological patterns during representation learning, thereby improving robustness to modality-specific variability~~in AI models addresses this limitation by training the models directly on suitable medical datasets, thereby enhancing the learning of clinically applicable features and improving robustness to modality-specific noise and variability~~. This review examines how different pretraining strategies influence the performance, robustness, and clinical reliability of AI models across various medical imaging modalities~~This study builds on the framework to evaluate how pre-training strategies guide model reliability across various imaging modalities~~.

This ~~manuscript~~ paper aims to evaluate ~~achieve~~ how domain-specific pre-training influences the clinical reliability, robustness and generalizability of AI models across four major medical imaging modalities: CXR, CT, MRI, and ultrasound~~some major imaging modalities which includes - CXR, CT, MRI and Ultrasound~~. Furthermore, it ~~also~~ critically compares domain-specific and general-purpose pre-training strategies, reviews the mechanisms underlying their performance differences and hence evaluates their implications for the ~~better~~ development of clinically reliable AI system in medical imaging.¶

Model architectures and the pre-training mechanisms in medical imaging

In the context of ~~a~~ this narrative review, mathematical or architectural derivation in detail is beyond the scope. The key concept is to properly understand ~~that~~ how different model architectures interact with pre-training strategies and how this interaction influences the performance across medical imaging tasks. As far as capturing spatial hierarchies, contextual information, and modality-specific patterns, these architectures differ.

Figure 1 illustrates a clear conceptual distinction between general-purpose and domain-specific pre-training strategies. While the general-purpose models transfer representations learned from the natural image datasets, domain-specific approaches learn modality relevant features (anatomical and pathological) directly from the medical imaging data, which improves both robustness and clinical reliability to a great extent.



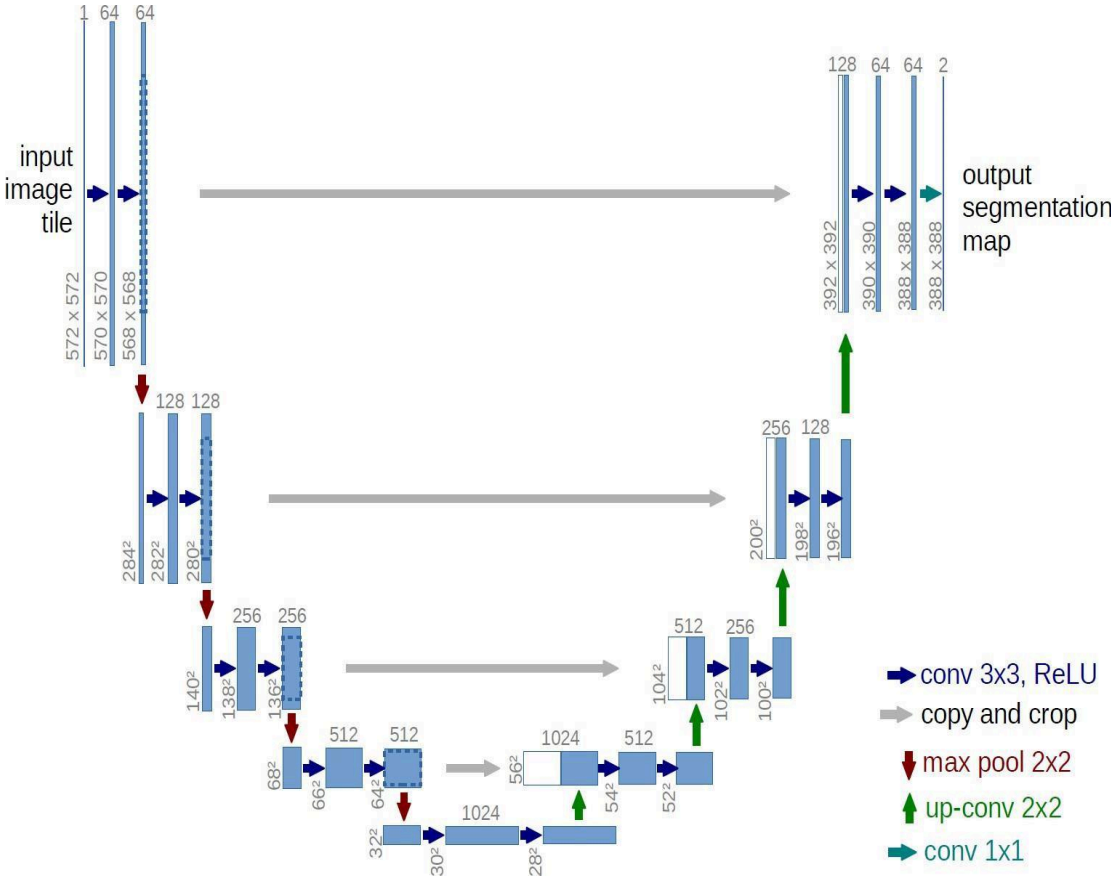
~~*Figure 1*~~¶

¶

*(**Figure 1—**. Conceptual comparison of general-purpose and domain-specific pre-training pipelines in medical imaging. General-purpose pre-training transfers representations which are learned from the natural image datasets whereas the domain-specific pre-training learns representations directly from the medical imaging data before task-specific fine-tuning.)*¶

¶

Encoder-decoder architectures such as U-Net have become very important for medical image segmentation, which involves identifying and precisely delineating anatomical structures or pathological regions within medical images. This is because they combine contextual feature extraction with precise spatial localization. As shown in Figure 2, skip connections preserve fine-grained anatomical information while integrating higher-level semantic features, thereby making U-Net highly effective for biomedical segmentation tasks (Ronneberger et al., 2015~~Encoder-decoder architectures such as U-Net have in particular becomebecome very important for medical image segmentation, a [define was imagine segmentation is]. This is because they combine contextual feature extraction with precise spatial localization. As shown in Figure 2, skip connections preserve fine-grained anatomical information while integrating higher-level semantic features, thereby making U-Net highly effective for biomedical segmentation tasks (Ronneberger et al., 2015~~).



~~Figure 2¶~~

~~(~~***Figure 2—***~~*. The U-Net architecture, an encoder–decoder network designed for biomedical image segmentation, in which the contracting path captures contextual information and the expanding path enables precise localization through skip connections (Ronneberger et al., 2015)~~*.*

During the process of pre-training, model optimization typically reduces a classification loss function such as cross-entropy loss through gradient descent-based learning (Goodfellow et al., 2016; LeCun et al., 2015). The learned weights steadily encode hierarchical feature representations, beginning with low-level spatial filters in earlier convolutional layers and slowly moving towards higher level semantic abstractions in the deeper layers (Zeiler & Fergus, 2014; LeCun et al., 2015). For classification tasks, optimization is commonly performed using the categorical cross entropy loss functions, which measures the differences between the predicted probabilities and the ground truth levels.

In medical imaging, the transferability of these learned representations depends strongly upon the resemblance between the source and target domains (Tajbakhsh et al., 2016; Raghu et al., 2019). When pre-trained on natural image datasets, optimization trajectories may become biased toward non-clinical visual features, limiting the ability of the model to identify pathology-specific structures (Raghu et al., 2019; Tajbakhsh et al., 2016). This representation mismatch becomes especially important in volumetric imaging modalities such as CT and MRI, where the contextual continuity across slices and subtle tissue-intensity relationships are clinically meaningful. Domain-specific pre-training somewhat addresses this issue by disclosing the optimization process to the medically relevant spatial and anatomical patterns during representation learning (Azizi et al., 2021; Raghu et al., 2019).

Consequently, the latent feature space learned by the model becomes more parallel with the clinically meaningful variations instead of generic object-recognition features. On the other hand, the domain-specific pre-training enables models to learn the feature representations which are directly aligned with medical imaging features. For example, 3D

CNNs which are trained on volumetric CT data can properly capture inter-slice spatial continuity that is clinically very important for anatomical interpretation (Çiçek et al., 2016). Similarly, MRI-focused models can learn relationships across multiple imaging sequences, thereby improving tissue characterization (Isensee et al., 2021~~Similarly, the MRI-focused models can learn relationships across multiple imaging sequences, including T1-weighted, T2-weighted, and contrast-enhanced acquisitions, thereby improving the tissue characterization (Isensee et al., 2021~~). Ultrasound-specific models have also been shown to improve robustness to speckle ¶

¶

noises and operator-dependent variability when trained on domain relevant datasets (Liu et al., 2019).

The effectiveness of these architectures dependsdepend strongly upon the alignment between the training data and the target clinical task. The case where the pre-training data closely matches the modality and task distribution, the models tend to perform well. This allows them to learn meaningful representations. Conversely, when there is a mismatch in the domain, the performance degrades which leads to poor generalization, increased false positives and unreliable clinical predictions. Therefore, it is the combination of architecture design and domain-specific pre-training that largely determines the success of AI systems in the field of medical imaging rather than the architecture alone (Raghu et al., 2019; Tajbakhsh et al., 2016; Azizi et al., 2021).

The advantages of domain-specific pre-training are also influenced by the dataset size and task complexity. Raghu et al. (2019) showed that ImageNet pre-training may provide modest improvements when medical datasets are small, but the advantage decreases gradually as task-specific medical datasets become larger. This indicates that the transfer learning effectiveness not only depends on architecture selection, but also on the dataset scale, modality complexity, and feature similarity between source and target domains.

Methodology

Data sources - This study is a narrative review that synthesizes existing literature on the role of domain-specific pre-training of AI models in medical imaging. Relevant studies were identified through the different searches conducted on databases such asincluding PubMed, IEEE Xplore, and arXiv.

Search strategy - Keywords used were used including, “mMedical iImaging”, “tTransfer lLearning”, “dDomain-sSpecific pPre-tTraining”, “MRI dDeep lLearning”, “CT sSegmentation”, “uUltrasound AI”, and “cChest xX-ray dDeep Learning”. Inclusion criteria included the peer-reviewed publications and some widely cited research papers that provided empirical evaluation of AI models in various medical imaging tasks. Widelylearning”. Widely cited papers were identified through proper database citation metrics and the through thorough examination of reference lists from highly relevant review articles and foundational studies in medical imaging AI. Full-text versions of potentially relevant studies were reviewed manually reviewed by the author to assess the methodological qualitymethodological clarity, empirical validation, and relevance to the objectives of this review as well.

Inclusion and exclusion criteria – Peer-reviewed studies and widely cited research papers that provided empirical evaluation of AI models in medical imaging were prioritized. Included studies directly compared domain-specific pre-training with general-purpose pre-training under comparable experimental settings, including image classification, segmentation, and detection tasks. Widely cited papers were selected when they represented foundational contributions to the field or were frequently referenced in subsequent medical imaging literaturePeer-reviewed sStudies which compared domain-specific pre-training with general-purpose pre-training were prioritized. Prioritized studies were the onesthose that directly compared domain-specific pre-trained AI models with general-purpose pre-training under comparable experimental settings, including image classification, segmentation, and detection tasks. Studies that mainly discussed theoretical concepts without proper empirical validation or did not clearly describe the datasets, evaluation metrics, or model training procedures were excluded from detailed analysis. Studies which compared domain-specific pre-training with general-purpose pre-training were included and discussed on the basis of modality-specific challenges throughout in depth in this paper. Methodological clarity was assessed based on on the basis of the availability of information regarding datasets, model architectures, evaluation protocols, and the reported performance metrics. Studies published between 2015 and 2023 were selected to reflect recent developments in deep learningDL.

Study scope and analysis - Studies which were published between 2015 and 2023 were prioritized to reflect recent developments in deep learning. This review is qualitative in nature and does not perform a formal meta-analysis. Instead, performance metrics which are reported in the original studies were are interpreted within the context of their datasets and the experimental conditions. Area Under the Receiver Operating Characteristic Curve (AUROC) was used to assess the discriminative performance in classification tasks because it measures an AI model's ability to distinguish between positive and negative cases across different decision thresholds. However, AUROC alone may not be sufficient for direct cross-study comparisons because datasets, disease prevalence, class distributions, and evaluation protocols can differ substantially between studies. The Area Under the Receiver Operating Characteristic Curve (AUROC) was used to assess the discrimination performances in classification tasks, while the Dice Similarity Coefficient (DSC) was used to evaluate overlap accuracy in segmentation tasks. Additional metrics such as sensitivity and specificity were considered

where reported. Performance comparisons across studies were interpreted cautiously because differences in datasets, class distributions, disease prevalence, evaluation protocols, and experimental settings can substantially influence reported performance metrics, including AUROC and DSC. MRI sequence terminology is also discussed throughout this review, where T1-weighted images primarily provide anatomical detail, T2-weighted images highlight fluid-containing structures and pathology and the contrast-enhanced T1-weighted (T1ce) images improve visualization of lesions and tumor boundaries (Bitar et al., 2006).

Results

The findings across the reviewed studies generally indicated that domain-specific pre-training was associated with ~~consistently~~ improved performance, robustness, and clinical applicability of the AI models across different ~~the various~~ imaging modalities. Table 1 summarizes the key imaging modalities, associated AI tasks, and their corresponding challenges. Tables 2–5 further compare representative studies across CXR, CT, MRI, and ultrasound imaging modalities, highlighting the differences in robustness, sensitivity, and generalizability between ~~the~~ domain-specific and ~~the~~ general-purpose pre-trained AI models.

Table 1: Overview of various ~~Imaging~~ imaging ~~Modalities~~modalities

Modality	Key AI tasks	Example domain-specific AI models	Challenges addressed
CXR	Pneumonia detection, multi-disease classification	CheXNet	Overlapping anatomical structures, subtle radiographic findings
CT	Lesion detection, volumetric segmentation	3D CNN, 3D U-Net	Volumetric context, scanner variability
MRI	Tumor segmentation, tumor classification	nnU-Net ~~: no-new-Net~~ (self-configuring U-Net framework)	Multi-sequence data, tissue-contrast variability
Ultrasound	Organ segmentation, echocardiography analysis	Ultrasound-specific CNNs	Speckle noise, operator dependency

~~(Performance comparisons across studies should be interpreted cautiously, as variations in datasets, evaluation protocols, and class distributions significantly influence reported AUROC and DSC.~~ Abbreviations: AI, artificial intelligence; CXR, chest x-ray; CT, computed tomography; 3D, 3-dimensional; CNN, convolutional neural network; MRI, magnetic resonance imaging; nnU-Net, no-new-Net). ~~(Performance comparisons across studies should be interpreted cautiously, as variations in datasets, evaluation protocols, and class distributions significantly influence reported AUROC and DSC).¶~~

Hardware workings of the various medical imaging modalities and their limitations

The CXR imaging is primarily based on the photons passing through the tissues of various densities present (Bushberg et al., 2012; Herring, 2020). While this method is fast and cost-effective, there are some limitations ~~to it as well~~, which ~~is that~~include the X-rays having many 2D projections with multiple overlapping anatomical structures, which limit the depth perception and also the sensitivity factors to subtle pathologies such as early COVID-19 infiltrates and other kinds of severe lung diseases (Jacobi et al., 2020; Herring, 2020).

CT, is widely used currently in the healthcare industry. The CT scans reconstruct multiple X-ray projections into volumetric 3-dimensional (3D) representations using various advanced tomographic reconstruction algorithms and therefore it offers superior resolution and also shows a proper and a better density variation (Bushberg et al., 2012; Kalender, 2011), but it also has a major limitation that is it exposes the patient to a higher amount of radiation doses which is extremely harmful (Brenner & Hall, 2007).

MRI relies completely on the nuclear magnetic resonance principles to capture the tissue-specific contrast through variations in the relaxation timings (T1, T2) and the pulse sequences (Brown & Semelka, 2010; Westbrook et al., 2018). MRI also provides an exceptionally soft tissue contrast but on the other hand it suffers from extremely high cost, long scan times and also scanner-dependent variability (Brown & Semelka, 2010; Westbrook et al., 2018).

Ultrasound imaging, uses high frequency sound waves to generate real-time images (Hoskins et al., 2010). The advantage of this is that it is safe, portable and cheaper compared to the other imaging modalities but it is highly sensitive to speckle noises and operator dependency (Hoskins et al., 2010; Noble & Boukerroui, 2006).

These modality-specific hardware characteristics directly influence the feature distributions available to the AI models, thereby affecting the effectiveness of different pre-training strategies and contributing to variations in model generalizability across the imaging modalities.

Chest X-Ray Imaging

One of the most widely used diagnostic modalities commonly used in clinical practices is the CXR. This is generally used for the detection and also monitoring of pulmonary and cardiovascular diseases. Since the cost is low, , the outcome is prompt, and it is also broadly available, chest radiology is considered to be the primary and the first imaging tool for detecting conditions such as pneumonia, tuberculosis, lung cancer, COVID-19, pleural effusion, heart failure and other such things.

¶

CXRs present multiple analytical challenges for both automated systems and AI models because the images are two-dimensional (2D), with overlapping anatomical structures and subtle intensity variations that often require expert interpretation. Consequently, CXR analysis remains a complex task, and several studies suggest that domain-specific pre-training enables AI models to learn radiographic features that are more relevant for clinical interpretation than representations learned from natural-image datasets (Raghu et al., 2019; Litjens et al., 2017CXRs present multiple analytical challenges for both automated systems and AI models. One of the most important reasons¶

¶

is that the images are two-dimensional (2D), and overlapping anatomical structures together with subtle intensity variations often require expert interpretations. Consequently, CXR analysis remains a complex task and several studies have reported that the domain-specific pre-trained models frequently outperform the general-purpose foundation models in clinically relevant tasks when they are evaluated under comparable conditions (Raghu et al., 2019; Litjens et al., 2017).

¶

Domain-specific pre-trained CXR models properly learn representations which are more closely aligned with the clinically relevant radiographic features, including lung field asymmetries, interstitial markings, and subtle opacities as well. Such features are often difficult to be identified using models that are pre-trained solely on natural image datasets. General-purpose AI models that are trained on datasets such as ImageNet may struggle to recognize subtle disease-specific patterns because they have a limited exposure to medically relevant imaging characteristics (Zech et al., 2018).¶

One of the most prominent examples of such training in the case of CXR imaging is *CheXNet,* proposed by Rajpurkar et. al. (2017). Beyond pneumonia, various recent studies and literature have demonstrated the effectiveness of the usage of domain-specific AI models in detecting COVID-19 related things, tuberculosis cavitation and lung nodules as well, which are often subtle and easily missed by the general-purpose AI models. Since it was trained on large-scale expertly labelled X-ray datasets, CheXNet achieved state-of-the-art performance on multiple thoracic diseases, including pneumonia, with AUROC scores typically in the low-to-mid 0.80s on the ChestX-ray14 test set, comparable to radiologist-level performance (Rajpurkar et al., 2017).

¶

AUROC is commonly used to evaluate the diagnostic classification performances because it measures an AI model's ability to distinguish between the positive and the negative cases across various different decision thresholds. However, AUROC alone is not always sufficient for cross-study comparisons because datasets, disease prevalence, class distributions and evaluation protocols can differ substantially between studies.¶

It was also found that CheXNet illustrated improved diagnostic performances in several reviewed parameters as compared to the general CNNs. On the other hand, the performances of general-purpose CNNs pre-trained on natural images were stated to show limited generalizability in certain clinical settings (Zech et al., 2018). Zech et al. (2018) also showed that the AI models that were trained on CXRs from a single institution were unable to generalize to external datasets. Especially when pretrained on non-medical images.

These findings suggest that the performance gaps between domain-specific and general-purpose pre-training are strongly associated with feature transferability and the dataset biases. These differences are likely related to the limited transferability of features learned from natural-image datasets to chest radiography, particularly when disease manifestations are subtle (Zech et al., 2018; Roberts et al., 2021)While ImageNet-pretrained CNNs may successfully learn generic edge and texture detections, they often fail to capture clinically relevant radiographic patterns when disease manifestations are subtle. -In chest radiography, models sometimes rely on false correlations such as hospital specific markers, scanner artefacts, or even acquisition patterns instead of the pathology-specific anatomical features (Zech et al., 2018; Roberts et al., 2021). This issue becomes very important especially during external validation, where the models trained under narrow institutional distributions may experience substantial drops in performances.

Collectively, these findings suggest that domain-specific pre-training may contribute to an improved diagnostic performance, robustness, and generalizability in CXR analysis when compared to general-purpose pre-training methods. However, the magnitude of these benefits may vary greatly depending on the dataset characteristics, model architecture, and evaluation settings. There are more recent studies that comparatively evaluated and found that the models pretrained on large-scale CXR datasets most of the times performed much better when compared to the general-purpose vision models of comparable sizes in both the classification and segmentation tasks (Irvin et al., 2019).

Domain-specific pre-training may also improve model interpretability by encouraging AI systems to focus on clinically relevant anatomical regions rather than non-diagnostic image artefacts. Several explainability studies using gradient-based localization techniques have shown that domain-trained AI models more consistently highlight pulmonary opacities, nodules, and other diagnostically relevant thoracic regions (Rajpurkar et al., 2017; Roberts et al., 2021AI models pre-trained on chest radiography datasets are generally more aligned with clinically relevant radiographic features and diagnostic patterns. Due to this the findings are better, and false positives are much less. This alignment is very important; since these are related to health, a small misclassification may lead to serious clinical consequences. The phrase “aligned with radiological reasoning” refers to the ability of the domain-specific foundational AI models to prioritize anatomically and clinically relevant regions throughout prediction rather than relying on non-diagnostic image artefacts or shortcut

~~correlations. Several explainability studies using relevant maps and gradient-based localization techniques have demonstrated that domain-trained AI models more consistently focused on pulmonary opacities, nodules, and clinically relevant thoracic regions compared to the AI models transferred directly from natural-image datasets (Rajpurkar et al., 2017; Roberts et al., 2021~~).

These findings indicate that both the foundational and the more recent studies consistently report potential benefits of using domain-specific pre-training for CXR analysis. Taken together, the evidence which is available suggests that the modality ~~-~~¶

¶

aware pre-training may play a crucial role in improving the clinical reliability of AI systems used in chest radiography. The importance of domain-specific pre-training is further supported by the availability of large-scale medical imaging datasets. For example, the CheXpert dataset contains 224,316 chest radiographs from 65,240 patients, providing substantial modality-specific data for representation learning and model development (Irvin et al., 2019). The ~~deep learning~~DL models which were trained on CXR images specifically achieved a higher multi-pathology classification performance when compared to the general foundation AI models.

Table 2 summarizes representative comparisons between CXR specific models and ImageNet-pretrained CNNs across pneumonia classification and multi-disease detection tasks.

Table 2: Chest X-Ray AI Model Comparison

Model	Pre-training	Task	Performance	Notes
CheXNet	CXR dataset	Pneumonia classification	AUROC ≈ 0.82–0.85	Evaluated on ChestX-ray14 dataset (Rajpurkar et al., 2017)
CNN	ImageNet	Pneumonia classification	Performance varies across datasets	Reduced external generalizability reported (Zech et al., 2018)
CheXNet	CXR dataset	Multi-disease detection	Strong performance across multiple thoracic diseases	Supported by multiple studies

~~(Table 2: Domain-specific models such as CheXNet demonstrate improved robustness and generalizability compared to general-purpose CNNs, particularly under cross-institutional evaluation settings.; Performance comparisons across studies should be interpreted cautiously, as variations in datasets, evaluation protocols, and class distributions significantly influence reported AUROC and DSC~~Abbreviations: CXR, chest x-ray; AUROC, Area Under the Receiver Operating Characteristic Curve; CNN, convolutional neural network~~)~~.

Computed Tomography Imaging

CT imaging plays a ~~very~~ vital role in day to day clinical workflows, especially in oncology and trauma assessment tests (Kalender, 2011; Bushberg et al., 2012). CT scan imaging is also widely used in the detection of pulmonary embolism, strokes, intracranial hemorrhages, and COVID-19 related abnormalities (Rubin et al., 2015; Kwee & Kwee, 2020). CT scans provide a high-resolution 3D volumetric data which properly capture finely grained anatomical structures and details across a wide range of slices (Kalender, 2011). These characteristics demand AI models that are capable of understanding complex 3D spatial relationships, tissue-density variations, and subtle pathological patterns (Çiçek et al., 2016; Litjens et al., 2017).

CT imaging represents one of the most complex modalities for reliable automated medical analysis. General purpose vision models trained on two dimensional (2D) natural images often struggle to interpret volumetric CT data effectively because they are designed primarily for planar image representations rather than volumetric feature learning (Raghu et al., 2019; Tajbakhsh et al., 2016). These models therefore lack intrinsic mechanisms for learning spatial continuity across adjacent slices and may thereby fail to properly capture clinically relevant 3D contextual information (Çiçek et al., 2016; Isensee et al., 2021).

In contrast, the AI models trained on volumetric CT data demonstrate highly improved sensitivity and robustness in the clinical detection and segmentation tasks (Çiçek et al., 2016; Setio et al., 2016). Previous studies demonstrate that CT specific models which are trained on volumetric data improve cancer detection sensitivity and lesion localization to a great extent when compared to 2D vision models pretrained on natural images (Çiçek et al., 2016; Setio et al., 2016). Several

studies show high improvements in sensitivity for lesion detection when CT-specific 3D architectures are used in comparison to 2D vision models (Setio et al., 2016).

CT-specific domain-trained models have been demonstrated to reduce false positives in lesion detection tasks, which is an¶
¶

extremely important and crucial factor in minimizing unnecessary follow-up procedures (Setio et al., 2016). These findings show the disadvantages of the general-purpose foundational AI models in handling volumetric medical data and also highlight the necessity of modality-aware pre-training.

A major difference between CT-specific architectures and conventional 2D CNN transfer learning approaches lies in their representations of spatial context. Standard ImageNet-pretrained CNNs process slices independently and therefore may lose inter-slice anatomical continuity, on the other hand, 3D CNNs and volumetric U-Net variants learn spatial dependencies directly across adjacent slices. These architectural differences become particularly important for detecting diffuse lesions, irregular tumor boundaries and low-contrast abnormalities that require volumetric contextual interpretation (Çiçek et al., 2016).

Another major advantage of CT-specific domain pre-training is in its ability to account for scanner-related variability and reconstruction artefacts. The CT images vary a lot, depending a lot on the scanner manufacturers, the radiation dose protocols and the reconstruction algorithms. The pre-training of AI models with specific domain data enables the AI models to learn invariant features properly and in depth so that they can generalize across such variations, whereas general-purpose vision models often wrongly interpret reconstruction artefacts as pathological features and hence limit their clinical reliability (Litjens et al., 2017; Roberts et al., 2021).

Domain-specific CT models are not completely without limitations. Although volumetric pre-training improves contextual understanding to a great extent, these models are computationally expensive and require mostly larger annotated datasets compared to the conventional 2D transfer learning approaches. Moreover, cross-institutional variability in CT acquisition protocols may still reduce the external generalizability despite domain-specific training. This suggests that architecture selection itself is not sufficient and must be combined with dataset diversity and thorough external validation.

Overall, the reviewed evidence highlights the limitations of general-purpose vision AI models for CT imaging and supports the use of domain-specific pre-training to improve performance, robustness, and clinical applicability. For complex applications such as cancer detection and trauma diagnosis, domain-specific CT AI models are more appropriate for achieving reliable clinical ~~Overall, this research and these studies demonstrate that CT imaging shows fundamental shortcomings in general-purpose foundational vision AI models, and it also strongly supports the adoption of domain-specific data pre-training methods. For complex and critical applications like cancer detection and trauma diagnosis, the domain-aware CT AI models are extremely needed for achieving a clinically reliable and trustworthy AI performance.~~performance.

Table 3 summarizes representative comparisons between CT-specific 3D architectures and conventional ImageNet-pretrained 2D CNN approaches in volumetric lesion detection tasks.

Table 3: Computed Tomography AI Performance

Model	Pre-training Type	Task	Performance	Notes
3D CNN	CT-specific dataset	Lesion detection	Higher sensitivity reported in controlled studies	Captures volumetric context (Setio et al. (2017))
2D CNN	ImageNet	Lesion detection	Lower sensitivity in volumetric tasks	Limited 3D understanding (Çiçek et al. (2016); Setio et al. (2017))

~~(Table3:~~ CT specific AI models such as CNNs achieve a higher performance in volumetric lesion detection tasks because they preserve volumetric special relationships. ~~Performance comparisons across studies should be interpreted cautiously, as variations in datasets, evaluation protocols, and class distributions significantly influence reported AUROC and DSC~~Abbreviations: 3D, 3-dimensional; CNN, convolutional neural network; CT, computed tomography; 2D, 2-dimensional~~)~~.

Magnetic Resonance Imaging

MRI has emerged as one of the most ~~important~~ widely used imaging tools in modern ~~-~~day medical diagnosis, especially in the fields of neurological imaging, oncology, and musculoskeletal imaging. Unlike other medical imaging tools such as X-ray technology-based medical imaging modalities, MRI image has very high-resolution volumetric data with very subtle contrast in soft tissues that are demonstrated through appropriate image acquisition protocols such as the T1-weighted image, the T2-weighted image, and the contrast-acquired image (T1ce image). This very multi-sequence and high-dimensional nature of the MRI data makes it in particular less acceptable for direct transfer learning from the natural image datasets.¶

The analysis of an MRI image becomes more complex for a general AI system, this is because there is a large variation in the intensities specified to MRI image capturing tools which is again combined with a large amount of anatomical details which were absent in natural images. Due to this, MRI serves as a critical platform for evaluation of the effectiveness of domain-specific pre-training in medical imaging models. The transfer learning from the natural image datasets most of the times provide a slight benefit in capturing the modality-specific intensity variations which are inherent to MRI data (Raghu et al., 2019).

A growing body of the existing literature continuously shows that the pre-trained AI models, those that are pre-trained on specific MRI datasets are much better than the general-purpose foundational AI models of same sizes across a wide range of tasks including tumor detection, classification and segmentation. Raghu et al. (2019) ~~elaborately~~ investigated the effectiveness of features learnt from the datasets which contained natural images in medical imaging tasks and demonstrated that the ImageNet pre-training provided very limited benefits for applications which were based on MRI.¶
¶
Their findings showed that when the CNN architectures of the similar sizes and capacities were compared, AI models pre-trained directly on MRI datasets consistently outperformed ImageNet-pretrained counterparts across medical imaging tasks such as brain tumor detection and tumor segmentation (Raghu et al., 2019). This limitation arises because MRI data contains modality-specific features, which includes tissue contrast mechanisms, multi-sequence information, and complex anatomical structures which are not present in the natural-image datasets. Consequently, representations learned from natural images may not adequately capture clinically relevant MRI characteristics, thereby limiting the effectiveness of direct transfer learning without domain-specific exposure (Raghu et al., 2019; Tajbakhsh et al., 2016; Litjens et al., 2017).¶
¶
These findings regarding the limited transferability of the natural-image pre-training are further supported by studies which demonstrate that the AI models which are trained directly on MRI data achieve superior tumor segmentation performances across multiple MRI sequences, including T1-weighted, T2-weighted, and contrast-enhanced T1-weighted (T1ce) images, when compared to the AI models initialized using natural-image datasets (Bakas et al., 2018; Isensee et al., 2021).¶
¶
This advantage in the performance has been continuously observed on public brain tumor segmentation benchmarks such as the Brain Tumor Segmentation (BraTS) Challenge (Bakas et al., 2018). Across BraTS benchmark evaluations, MRI-specific training strategies such as nnU-Net have consistently reported Dice similarity coefficients exceeding 0.80 for several tumor sub-regions, depending on the segmentation task and validation protocol (Isensee et al., 2021). On the contrary, evidence from transfer-learning studies suggests that ImageNet pre-training provides only a limited number of benefits for MRI-based segmentation tasks because the learned representations do not properly capture the MRI-specific tissue characteristics and anatomical structures (Raghu et al., 2019). These findings therefore highlight the limitations of relying solely on natural-image pre-training for MRI segmentation applications.

Cross-study comparisons show that the benefits of MRI-specific pre-training can become more pronounced as task complexity increases. While ImageNet transfer learning may provide moderate initialization advantages for simpler classification tasks with limited datasets, segmentation tasks which involve heterogeneous tumor boundaries and multi-sequence fusion will generally require modality-specific representational learning. This suggests that the effectiveness of transfer learning depends greatly upon the interaction between the task complexity, dataset scale, and modality-specific structural variation (Raghu et al., 2019; Isensee et al., 2021).¶
¶
These differences ~~in the performances~~in performance are particularly significant in clinical matters, where proper and precise tumor boundary demarcation directly influences the treatment planning and the diagnostic outcomes. In addition to these segmentation tasks, the MRI-specific foundational AI models have also demonstrated a much better performance in classification tasks like tumor grading and the detection of diseases. These tasks include stroke outcome prediction as well.

Multiple studies and literature have reported that the MRI-specific representational learning improves the disease classification and also the tumor grading performances when compared to the general-purpose foundational vision AI models, especially when the model sizes are the same (Raghu et al., 2019; Litjens et al., 2017). These improvements show the importance of AI models learning MRI-specific anatomical priors, which the general vision AI models often fail to capture effectively and properly. Other important and crucial factors influencing the MRI performances are scanner variability and acquisition heterogeneity. MRI images vary greatly across the scanner manufacturers, field strengths and patient demographics. Domain-specific pre-training with proper and wide MRI datasets exposes the AI models to the scanner and acquisition variability during learning, which leads to improved robustness and cross-institutional generalization, in contrast, the general-purpose models often fail to adapt reliably and effectively to the unseen MRI

distributions (Raghu et al., 2019;¶
¶

Roberts et al., 2021). Such robustness is extremely crucial for the safety and the reliability factor of clinical deployment of the MRI based AI models.

Some studies also indicate that domain-specific pre-training itself does not completely eliminate generalization failures. The differences in MRI acquisition protocols, magnetic field strengths, and institutional imaging practices may still produce significant performance variability during external validation. Therefore, although MRI-specific pre-training improves representation quality, robust clinical deployment additionally requires heterogeneous datasets, standardized evaluation protocols and rigorous external benchmarking as well (Roberts et al., 2021).

Overall, the extensive empirical evidence confirms that the MRI-based tasks substantially benefit from domain-specific pre-training with MRI-specific data when the AI models of the similar sizes are compared. These findings properly highlight that the MRI's modality-specific complexity requires proper and appropriate learning strategies and that the reliable clinical deployment of AI in MRI cannot rely solely on the general-purpose vision pre-training.

Table 4 summarizes representative comparisons between MRI-specific architectures and ImageNet-pretrained CNN approaches across segmentation tasks on BRATS and related benchmarks.

Table 4: Magnetic Resonance Imaging AI Performance

Model	Pre-training Type	Task	DSC / AUROC	Notes
nnU-Net	MRI dataset	Brain tumor segmentation	DSC ≈ 0.80+ to 0.90+	~~BRATS~~ BRaTS benchmark (Isensee et al., 2021)
CNN	ImageNet	Brain tumor segmentation	Typically, lower than MRI specific models	Limited transferability (Raghu et al. (2019))

~~(Table 4:~~ MRI specific AI models have frequently been reported to achieve a higher performance than ImageNet pretrained AI models because they learn modality specific tissue contrast and multi-sequence dependencies. ~~Performance comparisons across studies should be interpreted cautiously, as variations in datasets, evaluation protocols, and class distributions significantly influence reported AUROC and DSC.~~ Abbreviations: DSC, Dice Similarity Coefficient; AUROC, Area Under the Receiver Operating Characteristic Curve~~XYZ~~; nnU-Net, no-new-Net; MRI, Magnetic Resonance Imaging; BRaTS, Brain Tumor Segmentation Challenge~~XYZ DSC, Dice Similarity Coefficient~~; CNN, convolutional neural network~~)~~.

Ultrasound Imaging

Ultrasound imaging, due to its non-invasiveness, real-time acquisition and low cost, is being widely adopted across a wide range of clinical domains (e.g., abdominal imaging, obstetrics and cardiology). Other important applications include vascular assessment and echocardiography assessments as well. However, ultrasound imaging carries a number of complex and unique challenges, e.g., low signal-to-noise ratio, speckle noises, operator dependency, and considerable variability in image qualities. The general foundational vision AI models which are trained on natural image datasets may misinterpret speckle noises and acquisition artefacts as meaningful features, thereby leading to unreliable predictions (Xie et al., 2018; Chen et al., 2019).¶

¶

On the other hand, domain-specific pre-training on large-scale ultrasound data enables AI models to learn robust clinically relevant patterns despite the substantial noise and operator variability (Xie et al., 2018; Chen et al., 2019). Multiple surveys in the literature have shown that ~~deep learning~~DL models pre-trained on domain-specific ultrasound data consistently yield strong performance compared to general foundational AI models with comparable parameter sizes (Xie et al., 2018; Chen et al., 2019).

Performance metrics which include accuracy (the proportion of correctly classified cases), F1 score (the harmonic mean of precision and recall) and robustness under noisy imaging conditions generally favor the domain-trained ultrasound AI

models. To be specific, under noisy and operator-dependent imaging conditions, AI models pre-trained on domain-specific ultrasound data most of the times outperform general-purpose foundational AI models of comparable size (Xie et al., 2018; Chen et al., 2019).

Unlike CT and MRI, ultrasound imaging introduces a significant operator-dependent variability during image acquisition. Consequently, performance differences between domain-specific pre-training and general-purpose pre-training are not only influenced by representation learning, but also by acquisition inconsistency. Domain-specific ultrasound models are¶
¶

generally considered better at learning invariant features under noisy imaging conditions, while natural-image pretrained models may incorrectly amplify speckle artefacts or acquisition distortions during feature extraction. These differences are very critical when AI is taken and applied under real-world clinical settings, especially when ultrasound images vary significantly across devices and patient populations. This significant performance gap between general-purpose foundational vision AI models and ultrasound-specific AI models accentuates the need for modality-aware training in AI models in the healthcare field for trustworthy, accurate, and reliable results.

The literature surrounding ultrasound AI remains comparatively less standardized than MRI and CT research. Dataset sizes are often smaller, annotation quality varies substantially, and benchmarking protocols are less uniform across the studies. As a result, although current evidence strongly supports the advantages of ultrasound-specific pre-training, the direct comparison across studies remain difficult because of the differences in acquisition settings, devices, and operator expertise.

Table 5 summarizes representative comparisons between ultrasound-specific CNNs and ImageNet-pretrained CNN models across noisy and operator-variable clinical imaging conditions.

¶

Table 5: Ultrasound AI Performance

Model	Pre-training Type	Task	Performance	Notes
Ultrasound-specific CNN	Ultrasound dataset	Clinical tasks	Improved robustness under noise conditions	Domain adaptation beneficial (Xie et al. (2018); Chen et al. (2019))
CNN	ImageNet	Clinical tasks	Performance sensitive to noise and variability	Less robust (Xie et al. (2018); Chen et al. (2019))

~~(Table 5:~~ Studies reviewed in this paper suggest that domain-specific ultrasound AI models provide a greater robustness to noise and operator variability than general-purpose AI models~~. Performance comparisons across studies should be interpreted cautiously, as variations in datasets, evaluation protocols, and class distributions significantly influence reported AUROC and DSC~~. Abbreviations: CNN, convolutional neural network~~).

Discussion

Across the various imaging modalities ~~which is~~ are ~~discussed~~ in this review, the literature generally suggests that the AI models which are pre-trained on domain-specific medical datasets ~~most of the times demonstrate~~are associated with improved performance, robustness, and clinical applicability when compared to general-purpose foundational AI models (Tajbakhsh et al., 2016; Raghu et al., 2019; Zech et al., 2018; Roberts et al., 2021). Various findings from medical imaging studies suggest that modality specific details and pathologies are ~~most of the times~~often absent in the natural image datasets~~, due to which they may result in diminished performance~~ which may contribute to reduced performance in certain clinical tasks. Collectively, the reviewed studies indicate that the alignment between the pre-training dataset and the target imaging modality plays a vital role in determining feature transferability and downstream clinical performance ~~Therefore, we may conclude that the general-purpose foundational AI models rely on dataset specific patterns and shortcuts to a great extent in generalizing the practical applications in the healthcare sector~~ (Tajbakhsh et al., 2016; Raghu et al., 2019; Zech et al., 2018; Oakden-Rayner, 2020; Roberts et al., 2021). Sometimes, ~~the~~ general vision AI models may also have difficulty adapting ~~may find it complex in terms of variations~~to variations in image acquisition conditions, image reconstruction algorithms, and patient demographics as well. These variances are more effectively accommodated ~~mostly better handled through domain~~by domain-specific pre-training~~,s in~~ which may contribute to more reliable clinical predictions~~AI models due to which more reliable interpretations are developed~~.

The reviewed literature also indicates that the magnitude of improvement associated with domain-specific pre-training is not uniform across all imaging modalities~~A major pattern across the reviewed literature is that the benefits of~~

~~domain-specific pre-training are not even across all modalities~~. Simpler classification tasks with limited datasets might sometimes still benefit moderately from ImageNet initialization, especially when computational resources or annotated medical data are limited (Tajbakhsh et al., 2016). However, as task complexity increases, specifically in volumetric segmentation, multi-sequence analysis, or noisy imaging environments, the limitations of natural-image transfer learning become~~s~~ more apparent. Across the studies included in this review, CT and MRI applications generally reported larger improvements than CXR classification tasks because volumetric imaging and multi-sequence analysis require richer modality-specific feature representations. Similarly, ultrasound imaging benefited from domain-specific pre-training through improved robustness to speckle noise, operator dependency and acquisition variability. These findings suggest that the advantages of domain-specific pre-training becomes more pronounced as imaging complexity and contextual requirements increase~~This indicates that the impact of pre-training strategies depends strongly upon feature similarity, dataset scale, modality complexity, and external validation conditions as well~~.

The domain-specific image datasets have also shown considerable scope for further improvement~~great future scope to improve~~. There are widely used ~~a few largely used~~ image datasets that ~~in the domain-specific area which~~ are not fully expert-verified by radiologists, which may result in residual annotation variability ~~certified or verified in a correct manner by radiologists. Hence,~~¶

¶

~~residual variances may remain~~ (Rajpurkar et al., 2017). In addition, some image datasets also lack adequate demographic diversity, which may raise concerns regarding fairness and model generalizability ~~However, a few image datasets lack a variety of demographic populations due to which AI-related systems dedicated to healthcare may raise concerns in terms of fairness and reliability~~ (Oakden-Rayner, 2020; Roberts et al., 2021). To overcome these constraints, continued collaboration between clinicians, radiologists, and AI researchers is essential~~more interactions are needed between humans, doctors and AI researchers~~. Domain-specific AI models are not intended to replace doctors or radiologists. Instead, they serve as decision-support tools that can enhance ~~efficient tools to enhance~~ human-AI collaboration. When these AI models are developed using appropriately curated datasets and rigorous validation, they may function as a reliable second reader, helping to reduce radiologists' workload and support clinical decision-making. Nevertheless, the reviewed literature suggests that robust external validation and diverse, expert-annotated datasets remain essential for achieving clinically reliable AI systems~~When these models are correctly developed with accurate training and used practically, they can act as a reliable second reader. This can help relieve radiologists from heavy workloads and aid in sound clinical decision-making. However, using radiologist-verified datasets is crucial for improving fairness, safety, and reliability. The well-trained medical AI models can support radiologists by significantly reducing their workload and also acting as reliable readers. To understand the environments under which these AI models give clinically safe and trustworthy diagnostics results is extremely necessary for the application of AI in real world clinical practices~~. ¶

Though there are substantial advantages of domain-specific pre-training, there are certain limitations as well.

One major challenge is the limited availability of large-scale, high-quality annotated medical datasets, because expert annotation is both time-consuming and also very expensive. Moreover, many existing datasets lacks sufficient demographic diversity, which may lead to bias and limit the generalizability of AI models across different populations (Roberts et al., 2021; Oakden-Rayner, 2020). Furthermore, variability in imaging protocols, scanner types, and acquisition settings across institutions can affect model robustness despite domain-specific pre-training. The computational costs related to training large-scale domain-specific models also remain high, thereby posing practical constraints for widespread adoption.

When compared with alternative approaches such as transfer learning and emerging techniques such as meta-learning, domain-specific pre-training often provides better alignment between learned representations and modality-specific clinical features~~provides a stronger alignment between the learned representations and the modality-specific clinical features~~. Transfer learning enables adaptation of previously learned representations to new tasks using limited labelled data, whereas meta-learning aims to improve rapid generalization across multiple tasks and domains (Pan & Yang, 2010; Hospedales et al., 2021). However, several studies suggest that modality-specific pre-training may capture clinically better relevant imaging characteristics when sufficient domain data are available (Raghu et al., 2019; Tajbakhsh et al., 2016). Future research should therefore explore hybrid strategies that combine domain-specific representation learning with adaptable learning frameworks to improve both scalability and generalization (Hospedales et al., 2021).

Recent literature suggests that self-supervised learning may become an important complementary strategy in medical imaging. Unlike conventional supervised pre-training, self-supervised approaches learn representations directly from large volumes of unlabelled medical data through pretext or proxy tasks, thereby reducing dependence on expensive expert annotations (Azizi et al., 2021; Taleb et al., 2020). Several studies have reported that self-supervised domain-specific pre-training can improve robustness, feature transferability, and downstream task performance, particularly in data-limited clinical environments (Azizi et al., 2021; Taleb et al., 2020). However, external validation procedures and benchmarking frameworks remain insufficiently standardized across institutions, and further comparative research is required before these approaches can be integrated reliably into routine clinical workflows (Roberts et al., 2021).

Overall, the findings synthesized in this review suggests that the effectiveness of pre-training strategies depends on the interaction between imaging modality, dataset characteristics, task complexity and evaluation methodology. Rather than relying on a single universal pre-training strategy, future medical imaging systems are likely to benefit from modality-aware learning frameworks supported by diverse datasets, rigorous external validation, and clinically meaningful evaluation protocols.

Future Directions

Future research in medical imaging should focus not only on improving predictive performance but also on enhancing clinical reliability~~the predictive performances but it should also focus on improving clinical reliability~~, external validation, and explainability across multiple healthcare settings. One important direction involves the development of large-scale multi-institutional datasets that encompass ~~have~~ demographic diversity, heterogeneous scanner protocols, and geographically distributed patient populations. Such datasets ~~These kinds of datasets~~ would help reduce dataset-specific bias while improving ~~and also improve~~ model robustness during external validation (Roberts et al., 2021).

Another important research direction involves the development of multimodal learning frameworks that combine information from multiple imaging modalities, electronic health records, pathology reports and other relevant clinical metadata as well. Existing literature suggests that multimodal integration may improve contextual reasoning and reduce reliance on isolated image features, thereby potentially enhancing diagnostic consistency in complex clinical settings (Litjens et al., 2017~~Another major research direction involves multimodal learning frameworks that combine information from multiple imaging modalities, electronic health records, pathology reports, and clinical metadata as well. Existing literature suggests that multimodal integration may lead to better contextual reasoning and reduced dependence on isolated image features, therefore resulting in potentially improved diagnostic consistency in complex clinical settings (Litjens et al., 2017~~).

Future work should also prioritize explainability and uncertainty estimation in clinical AI systems. Although high AUROC or DSC values may indicate strong benchmark performance~~performances~~, these metrics alone~~benchmarks themselves~~ do not guarantee clinically reliable decision-making. Explainable AI techniques~~**frameworks**~~, saliency mapping methods~~techniques~~, and uncertainty -aware prediction frameworks may assist clinicians in interpreting model outputs more effectively while identifying predictions that require additional clinical review~~systems can help clinicians and doctors to better interpret model outputs and properly identify the high-risk prediction failures~~.

Greater systematic comparisons are ~~required~~ also needed between domain-specific pre-training, transfer learning, meta-learning, and self-supervised learning ~~approaches~~ under comparable evaluation conditions. Current literature often evaluates these approaches on different datasets and experimental settings, ~~which results in~~ making direct comparison more difficult. Benchmarking frameworks that incorporate common ~~which use consistent~~ datasets, external validation protocols, and clinically meaningful evaluation metrics would thereby improve reproducibility facilitate more reliable assessment of these learning strategies.~~and translational reliability.~~¶

¶

Finally, ~~In conclusion,~~ future clinical deployment research ~~researches~~ should prioritize prospective validation in real-world clinical environments rather than relying primarily ~~should prioritize prospective validation under real-world healthcare settings instead of relying mainly~~ on retrospective benchmark datasets. Several studies have shown that AI systems performing strongly in controlled experimental settings may sometimes fail under real clinical variability because of distribution shifts, demographic imbalance, or institutional differences (Roberts et al., 2021). Consequently~~Due to this~~, clinically deployable AI systems are likely to require continuous monitoring, recalibration, and collaborative oversight involving clinicians, radiologists and AI researchers to ensure sustained safety, reliability, and clinical effectiveness.

Conclusion

~~This review indicates suggests that d~~Domain-specific pre-training plays a vital role in improving the performance, robustness, and clinical applicability of AI models across various medical imaging modalities~~various different medical imaging modalities~~. ~~The reviewed literature indicates that t~~The effectiveness of pre-training strategies depends on factors including ~~various factors such as~~ modality characteristics, dataset scale, task complexity and external validation conditions as well. When compared to general-purpose vision AI models pre-trained ~~which are trained~~ on natural-image datasets, the domain-specific pre-training often enables the learning of representations that are more closely aligned with clinically relevant anatomical and pathological features.

The findings presented for ~~which are presented across~~ CXR, CT, MRI and ultrasound applications highlight the potential advantages of modality-aware learning strategies for supporting proper and reliable medical image analysis. However, clinically reliable ~~successful clinical deployment requires~~deployment requires more than just improved model performance ~~just improved AI model performance~~ alone. Diverse and representative datasets, rigorous external validation, standardized benchmarking frameworks, explainable AI methods and prospective real-world evaluation remain ~~highly~~ essential for improving ~~ensuring~~ safety and generalizability.

Future research should focus on multimodal data integration, cross-institutional validation, improved dataset diversity, systematic comparisons between domain-specific pre-training and alternative learning paradigms, and the development of clinically deployable AI systems that support effective human--AI collaboration as well. Taken together, these efforts may contribute to the development of more reliable, transparent and clinically useful AI systems for medical imaging.

Acknowledgement

The author of this paper acknowledges the contributions of all the researchers whose work formed the foundation of this review paper and also extends his sincere gratitude to his mentors and peers for their valuable guidance and feedback all throughout this research process.

Bibliography

1. 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.
2. Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems, 32.
3. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225.
4. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine, 15(11), e1002683.
5. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019, July). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 590-597).
6. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods, 18(2), 203-211.
7. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., ... & Jambawalikar, S. R. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629.

8. Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning?. IEEE transactions on medical imaging, 35(5), 1299-1312.
9. Setio, A. A. A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., Van Riel, S. J., ... & Van Ginneken, B. (2016). Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks. IEEE transactions on medical imaging, 35(5), 1160-1169.
10. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016, October). 3D U-Net: learning dense ¶
~~11.~~ ¶

~~12.~~ volumetric segmentation from sparse annotation. In International conference on medical image computing and computer-assisted intervention (pp. 424-432). Cham: Springer International Publishing.
13. Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S. X., ... & Wang, T. (2019). Deep learning in medical ultrasound analysis: a review. Engineering, 5(2), 261-275.
14. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC medicine, 17(1), 195.
15. Oakden-Rayner, L. (2020). Exploring large-scale public medical image datasets. Academic radiology, 27(1), 106-112.
16. Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., ... & Schönlieb, C. B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence, 3(3), 199-217.
17. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (pp. 234–241). Cham: Springer.
18. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., ... & Corrado, G. (2021). Big Self-Supervised Models Advance Medical Image Classification. Proceedings of ICCV 2021.¶
~~19.~~
20. Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2021). Meta-learning in neural networks: A survey. IEEE TPAMI.

Response to Action Editor

I sincerely thank the Action Editor for providing such a detailed and a constructive feedback on my manuscript. I greatly appreciate the careful review and the valuable suggestions regarding the Introduction, Methodology, Results, and reviewer response structure. These comments helped me improve the clarity, transparency, and academic rigor of the manuscript greatly.

Substantive Comments

Comment 1: There are now 2 aims towards the end of the Introduction (last and fourth-last paragraphs). Please consolidate these to only have one set of aims in the last paragraph.

Response 1: Thank you so much for identifying this issue. I have carefully revised the Introduction section and consolidated the duplicate aims into one final unified aim statement in the last paragraph of the Introduction. The earlier repeated aim statement was completely removed to improve clarity and flow. Relevant changes were properly made in the final portion of the "Introduction" section.

Comment 2: How were "some widely cited research papers that provided empirical evaluation of AI models in various medical imaging tasks" found? Was this by searching the reference lists of included papers, or were they searched in the databases sorted by citation numbers? This needs more clarification.

Response 2: Thank you so much for this important observation. I have properly expanded the Methodology section to clarify how the widely cited papers were identified. Specifically, I have added - citation metrics from databases were used, reference lists of foundational review articles; Influential studies were examined and manually screened full-text articles were reviewed for relevance and methodological quality. This clarification was then added to improve the transparency and reproducibility of the review process.

Comment 3: Who conducted the search? How were papers deemed relevant or not? This needs to be clearly articulated to show whether these decisions were made by one person or through some form of consensus.

Response 3: Thank you so much for this valuable suggestion. I have clarified within the Methodology section that the literature search, screening, and relevance assessment were conducted manually by the author through full-text review of the identified studies. I additionally clarified the criteria used to determine relevance, including empirical evaluation, methodological clarity, comparison of pre-training strategies and applicability to medical imaging tasks.

Comment 4: "Studies which compared domain-specific pre-training with general-purpose pre-training were prioritized..." What were these papers prioritised over?

Response 4: Thank you so much for pointing out this ambiguity. I have carefully revised this sentence in the Methodology section to clarify that studies directly comparing domain-specific and general-purpose pre-training under comparable experimental settings were prioritized over studies discussing theoretical concepts without empirical validation or detailed experimental methodology. This clarification was directly added into the revised methodology description.

Comment 5: What criteria were used to determine that a paper was to be excluded based on "Papers which lacked methodological clarity or empirical validation were excluded"? Who made these decisions, and what would constitute a lack of methodological clarity or empirical validation?

Response 5: Thank you so much for this important observation. I revised the Methodology section to explicitly define methodological clarity and empirical validation. The revised manuscript now explains that studies were excluded when they lacked - clearly described datasets, evaluation protocols, model architecture details. I have additionally clarified that these assessments were conducted manually through full-text review by the author (me).

Comment 6: Were full-text versions of papers reviewed, or was it just the Title and Abstract?

Response 6: Thank you so much for this clarification request. I have explicitly added to the Methodology section that the full-text versions of potentially relevant papers were manually reviewed by the author during the selection and evaluation process.

Comment 7: Tables 2–5 are still not mentioned in-text like Table 1 is.

Response 7: Thank you so much for identifying this issue. I have carefully revised the Results sections and explicitly incorporated references to - Table 2 in the Chest X-ray section, Table 3 in the CT section, Table 4 in the MRI section and Table 5 in the Ultrasound section. This was done to improve manuscript organization and consistency.

Comment 8: Please check abbreviations included under each table actually refer to the Table's contents.

Response 8: Thank you so much for this important formatting observation. I have carefully reviewed all table abbreviations and corrected inconsistencies throughout the manuscript. Abbreviations such as - AUROC, nnU-Net, CheXNet and US (Ultrasound) are now properly defined where necessary and redundant or unrelated abbreviations have been removed from table notes to ensure consistency with table contents.

Response to Managing Editor

I sincerely thank the Managing Editor for the thoughtful and such a highly constructive feedback provided on my manuscript. I greatly appreciate the highly detailed recommendations regarding the technical depth, comparative analysis, methodology clarification, and the academic writing quality. These comments significantly helped me to strengthen the manuscript and improve both its scholarly rigor and critical analysis. I have carefully revised the manuscript in accordance to all the suggestions.

Substantive Comments

Comment 1: (R1) There is still very little detail about how pretraining actually modifies representations or optimization, and I would appreciate a bit more explanation about the mathematics or model architecture.

Response 1: Thank you for this valuable suggestion. I have properly expanded the section titled "Model architectures and the pre-training mechanisms in medical imaging" to provide a deeper explanation regarding - representation learning, optimization behavior, hierarchical feature extraction and transferability limitations between natural-image and medical-image datasets. Specifically, I have added discussions regarding the following - hierarchical feature learning in CNNs, optimization using cross-entropy loss and gradient descent, representation mismatch between ImageNet and medical imaging and modality-specific feature learning in CT and MRI.

Comment 2: (R1) CNN/U-Net figures were added, but there are still no equations, training pipeline diagrams, feature-learning analysis, loss-function discussion, nor explanation of transfer mechanisms.

Response 2: Thank you for this valuable suggestion. To address this concern, I have added a detailed explanation regarding feature-learning mechanisms, transferability of learned representations, optimization behaviour, and the role of loss-function-based optimization during pre-training. I have also expanded the discussion regarding to the volumetric contextual learning in CT imaging and modality-specific representation learning in MRI (magnetic resonance imaging). As this research paper is a narrative review, I have carefully balanced the technical depth while maintaining readability and consistency with the scope of the review.

Comment 3: (R1) I do appreciate that the paper now gives some qualitative reasons for why models tend to succeed vs. fail, but can you do more? E.g. you can talk about feature transferability, data regime dependence, specific failure modes, model misrepresentation, etc. (maybe just pick a few of these to add)

Response 3: Thank you for this valuable suggestion. I have further expanded the manuscript by including more relevant critical analysis related to the following - feature transferability, dataset bias, shortcut learning, institutional overfitting, failure during external validation, modality-specific limitations and data-regime dependence. Additional comparisons discussing why ImageNet-pretrained models may fail in subtle pathology detection and volumetric imaging tasks compared to domain-specific models were added.

Comment 4: (R1) I still feel like the paper reads very summary-like rather than adding truly novel, critical analysis; a review paper goes far past just summarizing techniques and literature. Can you do some more direct comparisons between studies? Are there any limitations, conflicts, or contradictions in the literature? Can you compare and contrast some more learning techniques and competing approaches?

Response 4: Thank you for this valuable feedback. I have substantially revised multiple modality sections to properly include direct cross-study comparisons, limitations, contradictions, and modality-specific differences between learning approaches. The revised manuscript now compares domain-specific pre-training, ImageNet transfer learning, self-supervised learning and meta-learning approaches, while also discussing where transfer learning may still provide moderate benefits under limited-data conditions. This helped transform the manuscript from a primarily descriptive review into a more analytical and comparative scholarly review.

Comment 5: (R2) The performance metrics are now contextualized better, but studies are still mostly discussed sequentially rather than critically compared. There is a need for deeper cross-study analysis.

Response 5: Thank you for this valuable suggestion. I have revised the Chest X-ray, CT, MRI, and Ultrasound sections to include direct cross-study comparisons and analysis of external generalizability, task complexity, modality-specific differences, scanner variability and robustness limitations. Additional comparative discussion throughout the Results and Discussion sections was inserted.

Comment 6: (R2) You added some more future work to the conclusion, but could you make it a bit more actionable? Ideally, it should be in a separate section and well-developed by hearkening to the literature and showing where it wants to point, and you should be justifying why these are epistemically valid and actionable future directions.

Response 6: Thank you for valuable recommendation. I created a completely new section titled "Future Directions" before the

Conclusion section. This section now discusses - multimodal learning, explainable AI, self-supervised learning, external validation, uncertainty estimation, benchmarking frameworks, prospective clinical deployment and multi-institutional datasets as well. These future directions were properly supported using findings and concerns identified throughout the reviewed literature.

Comment 7: (R2) Some sections could be less repetitive and replace that repetitive content with deeper, critical analysis. Please carefully par down on the redundant, inflated wording (e.g., “critical, advanced and complex ones,” “reliable and trustworthy AI performance,” “proper and appropriate learning strategies,” etc.); more fluff and unnecessary verbosity are usually not good and prevent the reader from gleaning your more core points.

Response 7: Thank you for this valuable feedback. I have carefully revised all repetitive phrasings and removed inflated wordings throughout the manuscript. This has significantly improved conciseness and academic tone throughout the manuscript.

Comment 8: What do you mean by, “aligned with radiological reasoning?

Response 8: Thank you so much for identifying this ambiguity. I have clarified this statement within the Chest X-ray section by explicitly explaining that “alignment with radiological reasoning” refers to the ability of domain-specific models to focus on clinically relevant anatomical regions and pathology-specific structures rather than just relying on shortcut correlations or imaging artefacts.

- Be consistent with the use of "pre-training" vs. "pretraining" throughout the manuscript (choose one style and use it consistently)
- Be consistent with references to ImageNet pre-training (e.g., "ImageNet-pretrained," "ImageNet pre-trained," or "pretrained on ImageNet")
- Do a final check for subject-verb agreement. For example, "The effectiveness of architectures depend..." should be "depends..."
- If possible, reduce the repetitive use of phrases such as "clinical reliability," "general-purpose foundational AI models," and "domain-specific pre-training." They appear frequently and could occasionally be replaced with shorter wording or pronouns to improve readability.
- Ensure consistent hyphenation of compound adjectives throughout the manuscript (e.g., natural-image, domain-specific, modality-specific, task-specific, large-scale, cross-institutional, real-world, higher-level, low-level, day-to-day)
- In one of the "Abbreviations:" statements following a table, there is a punctuation error where CT is preceded by a semicolon instead of a comma