ARTICLE

Medical Data Mining 2026;9(2):10. https://doi.org/10.53388/MDM202609010

Brief application notes for vision transformer (ViT) and convolutional neural network (CNN) in medical imaging

*Correspondence to:

Author contributions

-

Competing interests

The authors declare no conflicts of interest.

Acknowledgments

-

Peer review information

Medical Data Mining thanks all anonymous reviewers for their contribution to the peer review of this paper

Abbreviations

Citation

Wong WK, Melinda M, et al. Brief application notes for vision transformer (ViT) and convolutional neural network (CNN) in medical imaging. *Med Data Min*. 2026;9(2):10. doi:10.53388/MDM202609010

Executive editor: Ming-Hao Wang.

Received: 20 September 2025; **Revised:** 17 November 2025

Accepted: ; **Available online:** 11 February 2026

Abstract

In contemporary computer vision, convolutional neural networks (CNNs) and vision transformers (ViTs) represent the two main approaches in image recognition. While both are used in various application in medical imaging, they work in fundamentally different ways. This report attempts to provide brief application notes on the ViTs and CNN in particular relating to situations to select one over the other. Generally, CNNs rely on convolutional kernels, local connections, and weight sharing. This gives them high efficiency and surprisingly strong performance in detecting features in local regions. ViTs, on the other hand, break images down into smaller sections (referred to as tokens) and use self-attention mechanisms to understand the relationships between all these sections, globally. In general, ViT achieves optimality when they are able to learn from incredibly large amounts of pre-training data. This report will examine briefly the structure, the underlying math, and the relative performance of CNNs and ViTs differ based on the lastest finding frmresearch work. Most importantly, the report serves as brief application note for implementation between the stated algorithms.

Keywords: convolutional neural network; vision transformer; comparative study; medical imaging

Introduction

Image processing has become an integral component in medical technology. The introduction of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) has ushered in an era where features are no longer human engineered, but self-generated within the framework of the model. While both CNNs and ViTs tackle similar problems specifically in medical imaging and object detection, they approach from dif ferent approach. CNNs rely on convolutional kernels, focusing on nearby connections and shared weights to create layered feature maps. ViTs, however, see an image as a string of patches, using self-attention to understand relationships across the whole picture. So, even though they both do well in the same areas, they need different amounts of data and computing power, and they have different built-in assumptions. As depicted in Figure 1, CNNs use a series of convolution and pooling steps, building up local features layer by layer, by implementing weight sharing. ViTs, by contrast, break the image down into patches, turn them into embeddings (adding positional info), and then use self-attention. This lets them "see" global connections right away. This difference in structure highlights why CNNs are great at picking up local textures, while ViTs are better at immediately understanding long-distance relationships.

In a vision transformer (ViT), an image is split into patches and each patch becomes a *token* (a vector). An input image of size $H \times W \times C$ (height, width, channels) is divided into fixed-size, non-overlapping patches of spatial size $P \times P$. Each patch is then flattened into a vector of length P^2C and linearly projected into an embedding of dimension d. This produces a sequence of tokens analogous to words in natural language processing.

In the original report introducing ViT [ref__199959], a 224 × 224 RGB image is split into patches of size 16×16. This yields 14 × 14 = 196 patches. Each patch has 16 × 16 × 3 = 768 raw values, which are projected to an embedding of dimension d = 768. The resulting 196 patch embeddings, along with an added classification token, form the input sequence to the Transformer encoder.

It is important to emphasize that patching works very differently as compared to convolution. Convolutions use overlapping sliding kernels with weight sharing, producing hierarchical local features. Patching, in contrast, is a rigid partitioning of theimage into non-overlapping tiles, each treated as a token. This design removes built-in inductive biases such as translation invariance, forcing the model to learn them from data.

As illustrated in Figure 2, the self-attention mechanism enables each image patch (token) to selectively integrate information from all other patches in the image. For an input token <wmath data-id="wmath-wej9a12cy" data-latex="$${{{x}}_i}$$" data-wrap="inline">xi</wmath>, three linear projections are computed: the query <wmath data-id="wmath-go97l0ywk" data-latex="$${{{Q}}_i}{=}{{{x}}_i}{{{W}}_Q}$$" data-wrap="inline">Qi=xiWQ</wmath>, the key <wmath data-id="wmath-7l8zdsl4q" data-latex="$${{{K}}_i}{=}{{{x}}_i}{{{W}}_K}$$" data-wrap="inline">Ki=xiWK</wmath>, and the value <wmath data-id="wmath-42ucwyr1p" data-latex="$${{{V}}_i}{=}{{{x}}_i}{{{W}}_V}$$" data-wrap="inline">Vi=xiWV</wmath>. The relevance between token i and any other token j is quantified using a scaled dot-product similarity as expressed in Equation 1.

<wmath data-id="wmath-w1qb4f5fc" data-latex="$${s_{ij}}{=}{{{{Q}}_i}{ }{{{K}}_j}}{{d}}{,}$$" data-wrap="inline">sij=Qi Kjd,</wmath> (1)

where d denotes the dimensionality of the key vectors. These similarity scores are normalised with a softmax function to produce attention weights as expressed in Equation 2.

<wmath data-id="wmath-9o2itynkx" data-latex="$${ _{ij}}{=}{exp({{s_{ij}}})}{_{k}^{}{exp}({{s_{ik}}})}{,}$$" data-wrap="inline"> ij=exp(sij) kexp(sik),</wmath> (2)

which represent the contribution of token j when updating token i. The output representation of token i is then obtained as a weighted aggregation of all value vectors as expressed in Equation 3.

<wmath data-id="wmath-h3jzjks1m" data-latex="$${{{z}}_i}{=}_{j}^{}{{ _{ij}}}{{{V}}_j}{.}$$" data-wrap="inline">zi= j ijVj.</wmath> (3)<wmath data-id="wmath-wh4qgoviz" data-latex="" data-wrap="block"></wmath>

This formulation allows global information exchange across the entire image, enabling Vision Transformers to model long-range spatial dependencies more effectively than convolution-based architectures.

Deep learning has become central to medical image analysis, with CNNs and ViTs emerging as the two dominant architectural paradigms. While both have demonstrated strong performance across a wide range of clinical applications, their relative strengths, limitations, and suitability for real-world medical deployment remain fragmented across the literature. This article presents a structured comparative analysis of CNNs and ViTs in medical imaging,focusing on key domains including radiology, ophthalmology, musculoskeletal imaging, and oncology. We examine how architectural inductive biases, data availability,pretraining strategies, and computational constraints influence model performance in tasks such as classification and segmentation.

Comparative analysis of CNNs and ViTs in medical applications

With the architectural foundations of CNNs and ViTs established, this section provides an analysis of their use across major medical imaging application domains. The section will briefly cover implementation in Radiology, Oncology, Dermatology/Histopathology, Ophthalmology, andMusculoskeletal Imaging. Most of these application involves either image segmentation or image classification. While the previous section explains with regard to image classification, slight modification enables both CNN and ViT to be used for pixel classification thereby creating a segmented region of interest.

Radiology: X-ray, CT, and MRI

Radiological imaging remains as one of the most extensively studied domains for comparing CNNs and ViTs. Across chest X-ray, CT, and MRI modalities, CNN-based approaches have demonstrated strong and reliable performance, particularly when annotated datasets are limited and diagnostically relevant cues are spatially localized. In COVID-19 detection from chest X-rays, CNNs trained either from scratch or via transfer learning achieved competitive accuracy, faster convergence, and greater training stability under label scarcity [2, 3]. Similar observations have been reported in brain MRI classification tasks, where CNNs maintained stable optimisation behaviour and efficient learning on moderately sized datasets [ref__199962].

However, radiological tasks often involve spatially distributed or diffuse patterns, such as lung opacities or infiltrative tumour growth, where purely local feature extraction may be insufficient. Vision transformers address this limitation through self-attention mechanisms that model long-range spatial dependencies. Multiple studies report that ViTs outperform CNNs in radiological tasks when adequate pretraining and fine-tuning are employed, particularly for complex MRI and CT analyses [ref__199963, ref__199964, ref__199965]. Murphy et al. further showed that pretrained ViTs exhibit improved transferability across radiology datasets, although their advantage diminishes in small-data regimes [ref__199961].

Hybrid CNN-ViT architectures have emerged as an effective compromise in radiology, particularly for segmentation. Ghribi et al. proposed a 3D U-Net-ViT hybrid for brain tumour segmentation [ref__199966], achieving a global accuracy of 99.56% and an average Dice similarity coefficient of 77.43%. These results highlight a recurring pattern: convolutional encoders provide precise local boundary and texture information, while transformer layers enhance global coherence across anatomically complex regions.

Oncological imaging and tumour detection

Oncological imaging deals with the detection, characterization and monitoring of tumours using medical imaging modalities including magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET) and histopathological whole-slide images. Imaging is an essential component throughout the cancer continuum of care, facilitating early diagnosis, treatment development, response evaluation and disease surveillance. Tumour morphology in medical images is highly heterogeneous and different in size, shape, texture and spatial arrangement between patients and types of cancer. Accordingly, automated image analysis is a growing area of study to ensure diagnostic rigor, lessen clinician load, and increase sensitivity to minor or precursory cancer. Oncological imaging is one of the most commonly used application domains to compare CNNs and ViTs, it can be implemented in classification or segmentation tasks. CNNs have been largely used for tumour

./image/file_697dc012dbc4b.png

./image/file_697dbf7d614e0.png

detection, attributed to the high capacity of CNN to represent lesion texture, edge area and regional intensity variation. However, results from different studies are not consistent. Some works show superiority of CNN in the fine-grained lesion analysis, while some work presented better performance of ViTs or hybrid architectures especially when tumours had heterogeneous or spatially distributed properties [ref__199967, ref__199968, ref__199969]. Figure 3 shows a representative MRI-based classification task, separating between tumour and non-tumour cases, as reported by Srinivasan et al. [ref__199970]. They present a hybrid CNN-ViT framework which merges convolutional feature extraction (CNN) and transformer-based attention, supporting the larger point that in oncological situations the ability to combine local and global representations is helpful. In the literature, the advantages of ViT in oncology are most readily observed when relying on pretraining and when the accuracy of segmentation relies on contextual relations between the tumour subregions rather than on purely local appearance.

EEG and MEG imaging

Electroencephalography (EEG) and magnetoencephalography (MEG) provide direct measurements of neural activity with millisecond temporal resolution, but their representation as multichannel time-series signals presents unique challenges for deep learning. Consequently, research in this domain has evolved from conventional CNN pipelines toward increasingly sophisticated hybrid and transformer-based architectures. Early and influential work demonstrated that CNNs can learn meaningful electrophysiological representations directly from raw or minimally processed EEG. For example, the compact EEGNet architecture established that carefully designed temporal and spatial convolutions can generalize across paradigms and subjects while remaining computationally efficient, explaining why CNNs remain dominant in low-data clinical settings [13, 14]. Subsequent work reinforced this finding, showing that CNNs can extract physiologically meaningful oscillatory and spatiotemporal features without handcrafted signal engineering [15, 16].

More recent studies increasingly report limitations of purely convolutional inductive bias, particularly for tasks where discriminative information is distributed across long temporal contexts or complex inter-channel relationships. Liu et al. proposed a hybrid CNN-Transformer architecture in which convolutional layers first learn stable local temporal-spectral features before transformer blocks model long-range dependencies across the signal, demonstrating improved classification accuracy compared with standalone CNNs on multiple EEG benchmarks [ref__199975]. Zhao et al. further extended this hybrid paradigm through the CNNViT-MILF framework, integrating CNNs, vision transformers, and multi-instance learning to explicitly address subject variability and weak labeling, two persistent challenges in real-world EEG analysis [ref__199976]. Their results suggest that attention mechanisms offer clear advantages in heterogeneous datasets where global relational modeling is essential.

Several recent ViT-centric studies provide direct evidence that attention-based architectures can outperform convolutional baselines when EEG is encoded as structured time-frequency or connectivity representations. Chakladar demonstrated that when EEG-derived brain connectivity matrices are treated as visual patterns, a vision transformer can more effectively capture global relational structure than CNNs, resulting in improved discrimination of neurological states [ref__199977]. Afzal et al. applied vision transformers to large-scale clinical seizure datasets and showed that transformerbased models generalize more robustly across patients than conventional convolutional approaches, emphasizing the importance of long-range temporal modeling for realistic clinical deployment [ref__199978]. Related work further supports this conclusion by demonstrating the benefits of multi-stream and multimodal ViT-based architectures for complex EEG conditions such as epilepsy and cognitive disorders [ref__199979, ref__199980, ref__199981, ref__199982].

MEG research exhibits a parallel evolution but with additional emphasis on spatial modeling and source inference. Wang et al. demonstrated that deep learning models, including CNNs and vision transformers, can approximate the traditionally ill-posed MEG inverse problem, enabling fast and accurate source localization directly from sensor data [ref__199983]. This result is significant because it suggests that data-driven models can replace computationally expensive biophysical solvers in time-critical MEG applications. Earlier work by Seeliger et al. already showed that CNNs can successfully decode and localize MEG activity patterns, reinforcing the suitability of convolutional architectures for structured spatiotemporal neural data [ref__199984].

More recent transformer-based approaches in MEG focus less on classification and more on foundational signal modeling tasks. Tibermacine et al. proposed attention-based models for MEG denoising, showing that transformer architectures can enhance event-related fields under low signal-to-noise conditions, thereby reducing the need for extensive trial averaging [ref__199985]. Khadka et al. further explored transformer-style architectures for long-sequence modeling of MEG time-series, demonstrating that attention mechanisms are well suited to capturing long-term temporal structure in continuous neural recordings [ref__199986]. Afzal et al. additionally argued for the role of large-scale transformer models as emerging foundation architectures for electrophysiological signals, positioning MEG and EEG within the broader trend toward pretraining and transferable neural representations [ref__199987].

Taken together, the literature consistently indicates that CNNs remain highly effective when data are limited and discriminative features are local, which explains their continued dominance in clinical EEG and MEG pipelines [ref__199971, ref__199972, ref__199973]. However, an increasing number of studies demonstrate that vision transformers and hybrid CNN-ViT architectures offer tangible advantages when tasks require modeling long-range temporal dependencies, distributed spatial interactions, cross-subject variability, or multimodal fusion [17-21, 25]. This trajectory mirrors broader trends in medical AI, where convolutional architectures provide strong baselines, but attention-based models increasingly define the frontier for large-scale, generalizable, and foundation-style learning.

Dermatology and histopathology

Dermatology and histopathology focus on the visual assessment of skin surfaces and microscopic tissue structures to diagnose inflammatory, infectious, and neoplastic diseases. Imaging modalities such as dermoscopic images and digitized histopathology slides are central to clinical decision-making, as diagnosis often relies on subtle textural patterns, boundary irregularities, and cellular morphology. These images are typically high-resolution and rich in fine-grained visual detail, making automated analysis particularly valuable for screening, workload reduction, and diagnostic standardization in both clinical and telemedicine settings.

In dermatology and histopathology, CNNs continue to provide strong baseline performance due to their locality bias, which aligns well with texture and edge dominated visual patterns such as skin lesions, cellular nuclei, and glandular structures. In dermoscopic image analysis, CNNs frequently achieve stable and competitive results even with limited annotations, making them suitable for screening and teledermatology applications [30, 31].

Vision transformers have also demonstrated the potential to match or exceed CNN performance in these domains, but predominantly under conditions of substantial pretraining and careful fine-tuning. Takahashi et al. showed that ViTs can achieve comparable classification accuracy in histopathological tasks [ref__199990], although at the cost of increased training complexity and computational demand. In digital pathology, where whole slide images contain both localized cellular features and broader tissue architecture, transformer-based models benefit from their ability to model long-range dependencies. Nevertheless, computational cost and sensitivity to patch selection strategies remain key barriers, reinforcing the appeal of hybrid CNN-ViT designs.

Ophthalmology and retinal imaging

Ophthalmology focuses on the diagnosis and management of diseases affecting the eye, many of which manifest as structural and vascular changes in the retina. Retinal imaging plays a central role in this field, as the retina provides a non-invasive window into ocular and systemic health. Imaging modalities such as fundus photography and optical coherence tomography (OCT) are routinely used to screen, diagnose, and monitor conditions including diabetic retinopathy, age-related macular degeneration, glaucoma, and hypertensive retinopathy. These diseases often progress silently in early stages, making automated retinal analysis particularly valuable for large-scale screening, early intervention, and resource-limited clinical settings, Ophthalmology is one of the earliest medical domains where deep learning systems have achieved regulatory-approved deployment. CNN-based models have been widely adopted for fundus photography and optical coherence

./image/file_697dc0984afb1.jpg

tomography (OCT), excelling at capturing vascular structures, optic disc morphology, and localized retinal lesions. Under limited labelled data, CNNs remain highly effective and computationally efficient.

Generally both ViT and CNN were often evaluated particularly for distinguishing between the various stages of Retinopathy or Glaukoma. The myriad of datasets (such as APTOS, Messidor) encourages the trend of utlising various deep learning approaches, [ref__199991] evaluated enhancing optimality in CNN transfer learning by using metaheuristic approaches and reported state of the art improvements. Vision transformers have shown advantages in retinal disease diagnosis when pathological patterns are spatially distributed across the field of view. Hwang et al. demonstrated that ViT-based models outperform CNNs after pretraining [ref__199992], particularly in multi-disease classification settings. Systematic reviews further indicate that attention mechanisms improve integration of spatially dispersed retinal cues and enhance cross-dataset robustness [ref__199969]. As with other domains, hybrid CNN-ViT architectures often offer the most practical balance between performance and data efficiency. The challenge lies in the fact that most of the research exclusively used 1 dataset for evaluation. Cross dataset evaluation could further show generalisation capability of the reported results.

Musculoskeletal imaging and bone assessment

Musculoskeletal imaging focuses on the assessment of bones, joints, and connective tissues for the diagnosis of degenerative, traumatic, and metabolic disorders. Conventional radiography remains the most widely used modality in this domain due to its low cost, accessibility, and suitability for population-level screening, particularly in the context of osteoporosis, osteoarthritis, fractures, and spinal deformities. Clinical interpretation commonly relies on subtle radiographic cues such as variations in bone density, cortical thickness, joint space narrowing, and overall skeletal morphology, which are inherently subjective and prone to inter-observer variability. Consequently, automated image analysis using deep learning has attracted substantial interest as a means of supporting early diagnosis, improving reproducibility, and enabling large-scale screening in ageing populations.

A substantial body of literature demonstrates the effectiveness of convolutional neural networks (CNNs) for musculoskeletal image analysis. Early studies showed that deep CNN models can reliably detect abnormalities in musculoskeletal radiographs and improve diagnostic efficiency in routine clinical workflows [35, 36]. Comprehensive reviews further confirm that CNN-based systems have achieved strong performance across a wide range of musculoskeletal tasks, including fracture detection, joint abnormality classification, and tissue segmentation in MRI, largely due to their ability to exploit local texture and structural patterns [ref__199995, ref__199996, ref__199997]. More recent contributions highlight that CNN-driven pipelines are increasingly being integrated into clinical practice as decision-support tools, particularly in high-volume radiology environments where workload and variability remain critical challenges [ref__199998]. These findings collectively support the continued relevance of convolutional architectures in musculoskeletal imaging, especially in settings where training data are moderate in size and diagnostically relevant features are spatially localized.

However, emerging evidence also indicates that purely convolutional inductive bias may be insufficient for tasks where diagnosis depends on more global structural relationships. In population-level screening applications such as osteoporosis assessment, the discriminative cues are often distributed across the entire skeletal structure rather than confined to isolated local regions. Cross-analysis studies suggest that when sufficient data or pretrained representations are available, transformer-based architectures can provide state-of-the-art performance by more effectively capturing long-range dependencies and holistic anatomical context. Bi et al. proposed a hybrid CNN-transformer architecture for musculoskeletal image analysis and demonstrated that incorporating attention mechanisms improves robustness and generalization compared with CNN-only baselines [ref__199999]. Similarly, recent studies applying global modeling strategies to musculoskeletal radiographs emphasize that structural relationships across distant anatomical regions contribute meaningfully to diagnostic accuracy, supporting the suitability of transformer-style representations for this domain [42, 43].

Within this context, Sarmadi et al. provide a direct comparative evaluation between CNN-based and vision transformer (ViT)-based models for bone assessment tasks [ref__200002]. Their results indicate that ViT-based architectures can significantly outperform convolutional baselines in specific musculoskeletal applications, particularly when diagnostic decisions depend on global anatomical structure rather than local texture alone. This finding aligns with broader trends observed across medical imaging, where CNNs remain highly effective under limited data regimes, but ViTs offer increasing advantages when tasks require holistic structural reasoning and when sufficient training data or transfer learning strategies are available.

A summary of the implementation discussion is shown in Table 1. It is observed that for most application, there are advocates for both ViT and CNN with justification.

Factors for selection between ViT and CNN

Previous section have discussed some of the works pertaining to the use of ViT and CNN. This section will attempt to evaluate/formalize the factors that should be considered for implementation. As discussed in previous section, CNNs continue to be a popular option when dealing with smaller datasets, costly annotations. Their inherent strengths lies in local connectivity and the ability to recognize patterns regardless of location. For instance, some studies comparing COVID-19 X-ray detection and brain MRI classification; authors noted that CNNs not only trained more quickly, but were also more reliable when dealing with limited labeled data [2, 4]. Further supporting this, research in dermatology and chest radiograph classification indicates CNNs offer consistent performance even with limited annotations. ViTs, on the other hand, often needed pre-training to achieve similar results [30, 45]. CNN efficiency is also crucial in safety-critical systems, where fast response times are essential [ref__200004]. Image classification involving smaller datasets, CNNs demonstrated better convergence and reduced the risk of overfitting compared to ViTs [47, 48]. Their reduced computational demands make them well-suited for edge devices and real-time applications, as seen in areas like industrial crack segmentation and geological fault detection [49, 50]. Further supporting this, it is noteworthy that areas such as dermatology and chest X-ray analysis, CNNs tend to deliver consistent results with minimal annotation; ViTs, on the other hand, may need pre-training to compete [30, 45]. Efficiency is also key in safety-critical setups, and studies underscore CNNs' low-latency advantages [ref__200004].

Shallower CNN layers tend to pick up the localized features (edges, textures), which is optimal for domains such as dermatology and histopathology. ViTs, in comparison, use global self-attention across all patches, regardless of length. As such CNNs can still dominate in situations where fine-grained, localized structures are what really matter. Conversely, ViTs utilize self-attention globally across all image patches from the outset, lacking that intrinsic sense of locality. It's this absence that likely explains CNNs' continued superiority in fields where subtle, local details hold the most crucial information.

Vision Transformers (ViTs) perform optimally when a task demands understanding across an entire image or even through time. Human action recognition, for instance. CNNs, being more focused on local motion, seem to miss global context [ref__200009]. This advantage in ViT is important especially in identifying tumors, ViTs leveraged on these global perspective to outperform CNNs on similar task as demonstrated by the various research discussed, but only with training of more data training data [ref__199963]. Reinforcing this, authors in other domain cited that ViT they were much better at understanding satellite images, thanks to their ability to connect broader regions of the image. CNNs struggled with those wide-area relationships as [52, 53]. Fault detection in geology followed a similar pattern: ViTs were more accurate by making sense of the overall picture in high-resolution images [ref__200008]. This trend holds up in medical imaging. Research on eye images showed that ViTs captured those widespread features associated with glaucoma better than CNNs [ref__199992]. And a review of different architectures suggested that ViTs are great at combining multimodal information such as images, clinical data,because they're flexible with that attention mechanism in the architecture [11, 54].

It's generally understood that Vision Transformers (ViTs) excells when there is large, annotated datasets or the opportunity for transfer learning. In these situations, ViTs have often proven themselves superior to Convolutional Neural Networks (CNNs). This is largely because their comparatively minimal inductive bias becomes relevant when working at larger scale, allowing for representations that are more adaptable. Some studies show ViTs achieving a slight improvement in accuracy over CNN baselines on benchmarks such as ImageNet and CIFAR [4, 55]. Report in(Tokens-to-Token ViT) reports that its T2T-ViT model achieves 1.4%-2.7% higher top-1 accuracy than ResNet-50/101/152 under comparable model size and computational cost on the ImageNet benchmark [ref__199962], demonstrating that transformer-based architectures can outperform established CNN baselines when training protocols are carefully controlled. Similarly, report in [ref__200013] (Comparative Analysis between CNN and ViT using Brain MRI Dataset) demonstrates that a Vision Transformer achieved 88.5% classification accuracy compared to 85.5% for a CNN model on a brain tumor MRI dataset, corresponding to a 3% absolute improvement in a real medical imaging task further Supporting this statement authors have shown through pigmented skin lesion classification. Here, pre-trained ViTs, after being fine-tuned on clinical images, actually surpassed the performance of CNNs [ref__199989].

ViTs, consistently outpace CNNs on large-scale benchmarks when pretraining is part of the consideration. Studies suggest that pretrained ViTs create flexible representations that generalize well across different tasks. CNNs trained from the ground up, on the other hand, often need a lot more fine-tuning [3, 6]. Consider radiology, for instance, where some pretrained ViTs demonstrated better sample efficiency and were more robust to hidden stratification (unlabeled subgroups within a class) than their CNN counterparts [ref__199961]. Pretrained ViTs have also been successfully applied to dental radiology, delivering competitive baseline performance across various downstream classification tasks [ref__200014].

While CNNs will continue to be relevant despite the advantages posed by ViT, particularly when training directly on smaller or medium-sized datasets, ViTs that have been trained on truly massive datasets and then fine-tuned tend to outperform CNNs. This is especially true in areas like medical imaging, security, and remote sensing. From the exiting reports, scale and pretraining enables the full potential of ViTs. Table 2 summarises the considerations for consiering between ViT and CNN.

From a mathematical perspective, error appears to follow a power-law with dataset size: <wmath data-id=”wmath-96pvynrms” data-latex=”$${ }{{ }}({N}){ }a{N^{{−} }}{+}b$$” data-wrap=”inline”> (N) aN− +b</wmath>. If pretrain , ViTs tend to show a larger exponent <wmath data-id=”wmath-g2yd7bq3j” data-latex=”$${ _{{ViT}}}{>}{ _{{CNN}}}$$” data-wrap=”inline”> ViT> CNN</wmath> and a lower asymptote b, which helps explain steeper performance improvements and higher limits when really scale up [ref__200015].

In response to the weaknesses and strengths of both, it is important to highlight that hybrid of ViT and CNN were often considered. For instance, hybrid models that explain themselves well are being used for crack segmentation, offering improved interpretability and robustness compared to CNN-only setups [ref__200007]. In situations such as federated learning, where data is spread out, hybrids can strike a good balance between efficient communication and powerful representation, using the CNN to compress local features and the ViT to flexibly combine information on a global scale [ref__199963].

Intuitively,in the hybrid architecture, CNNs can stabilize training and lessen the need for huge datasets, while transformer blocks add adaptability and scalability [59, 60]. All in all, hybrid architectures are generally seen as a “best of both worlds” solution. That makes them particularly attractive for real-world scenarios where you need both high accuracy and good efficiency.

Concluding remarks

This study provides brief comparative study between CNNs and ViTs, though often pitted against each other, each should be selected based on the application. CNNs, with their ingrained understanding of local patterns and lean designs, continue to be workhorses in situations where resources are tight, datasets are modest, and local details reign supreme. On the other hand, ViTs reaches full potential when afforded the opportunity to absorb global context and leverage the power of extensive pre-training.