International Journal of Applied Mathematics

Volume 38 No. 3s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

DEEP LEARNING FOR MEDICAL IMAGE ANALYSIS: A REVOLUTION IN DIAGNOSTIC MEDICINE

Ahmed A.F Osman [1,a*], Mohammed Awad Mohammed Ataelfadiel [2,b]

[1] Applied College, King Faisal University, P.O. Box 400, Al-Ahsa 31982, Saudi Arabia

[a] afadol@kfu.edu.sa;*

[2]Applied College, King Faisal University, P.O. Box 400, Al-Ahsa 31982, Saudi Arabia , [b] melfadiel@Kfu.edu.sa

**Corresponding author E-mail: afadol@kfu.edu.sa*

Abstract

Deep learning has revolutionized medical image analysis by enabling rapid, automated inter- pretation of complex imaging data. We propose a novel framework that combines a self-configuring U-Net backbone with integrated transformer-style attention modules to en- hance both local feature extraction and global contextual reasoning. Using publicly available datasets—the Synapse multi-organ CT challenge and the ACDC cardiac MRI collection—we demonstrate that our method achieves mean Dice scores of 0.83 and 0.77, respectively, with competitive 95th-percentile Hausdorff distances, all under CPU-only training conditions. Qualitative visualizations, per-class performance analysis, and volume-versus-accuracy stud- ies reveal robust segmentation of challenging anatomical structures. Comparative evaluation against CE-Net, TransUNet, and 3D TransUNet shows that our approach approaches state-of-the-art benchmarks without extensive manual tuning. Our framework's self-configuration and attention mechanisms offer a reproducible, resource-efficient solution for clinical deployment and pave the way for further enhancements through multi-modal fu- sion and domain adaptation.

Keywords: Deep Learning; Medical Image Segmentation; Self-configuring Networks; Attention Mechanisms; Transformer U-Net.

1. Introduction

Medical imaging is central to modern diagnostics and treatment planning. Radiology, cardiology, oncology, and ophthalmology, among many others, depend on imaging for the detection of disease, staging, and the monitoring of therapy. Traditionally, quantitative image analysis has relied on handcrafted features (intensity statistics, texture descriptors, shape features) that are often combined into "radiomics" pipelines with classical machine learning classifiers (such as support vector machines, random forests) [1]–[2]. These approaches enabled prognostic and predictive modeling but demanded expert feature engineering and were sensitive to choices of acquisition protocols and preprocessing [1]–[3].

The advent of deep learning, particularly deep convolutional neural networks (CNNs), fundamentally changed how visual information is represented and processed. In the general computer vision community, AlexNet's breakthrough ImageNet performance demonstrated the power of supervised deep CNNs for large-scale image classification [4], followed by very deep architectures such as VGG [5] and residual networks (ResNet) [6]. Parallel advances in optimisation, hardware acceleration, and open-source software further lowered the barrier to deploying deep models. A broad conceptual consolidation of these developments was articulated in LeCun, Bengio and Hinton's "Deep learning" review.[7]

Medical image analysis quickly adopted these advances. U-Net introduced an encoder–decoder architecture with skip connections tailored to biomedical image segmentation [8], which rapidly became the de facto backbone for many segmentation tasks. Generative adversarial networks unlocked new possibilities for image synthesis, domain adaptation, and reconstruction [8]. Comprehensive surveys, such as that of Litjens et al., providing an overview of deep learning in medical image analysis [9], and that of Lundervold and Lundervold with a review focused on MRI [10], documenting an explosion of work across modalities and tasks. More recent reviews have addressed radiomics and imaging AI more broadly, often with a strong emphasis on diagnostic performance and technical trends [1]–[3], [9], [10].

There is increasing recognition from a clinical perspective that AI has the potential to augment but not replace clinicians in high-throughput diagnostic workflows. Topol argued that the synergy of human expertise and AI systems will underpin “high-performance medicine” and stressed that this will require careful attention to validation, transparency, and workforce implications [11]. Kelly et al. identified salient barriers to clinical impact, including data quality, generalizability, regulatory constraints, and sociotechnical integration.[12]

1.1 Positioning relative to existing reviews

While previous surveys have indeed presented important overviews of deep learning in medical imaging [3], [9], [10], they have typically focused on either specific modality-e.g., MRI [10], or subspecialties, that of radiology-or predominantly on technical performance metrics. Many were also written prior to the emergence of transformers, large-scale self-supervised learning, and multimodal medical foundation models.

This review differs from and extends these works in three keyways:

1 .**Integration of foundational and emerging paradigms:** We bridge classical CNN-based architectures with more modern developments in transformers, selfsupervised learning, and multimodal foundation models by highlighting how these paradigms reshape opportunities and challenges in medical imaging [13]–[15][33].

International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

2 .**Dual organization by modality and clinical tasks:** We jointly structure the field along imaging modalities - e.g., radiography, CT, MRI, ultrasound, digital pathology, ophthalmic imaging-and clinical tasks - i.e., classification, detection, segmentation, reconstruction, prognosis-showcasing how architectural choices and evaluation strategies vary across this landscape.

3 .**Explicit focuses on clinical translation, fairness, and governance**: Beyond methodological advances, we foreground issues such as dataset shift, robustness, fairness, explainability, regulatory frameworks, and workflow integration, drawing on emerging guidelines such as CONSORT-AI, SPIRIT-AI, and STARD-AI [11], [12], [34]–[37][40].

Our goal is to provide both technical and clinical readers with a coherent narrative that links algorithmic progress to real world diagnostic medicine.

2. Methods

Although this paper is not systematic but rather a narrative review study, we applied a structured process to find eligible studies for citation within this paper with the point of ensuring recency and relevance of studies.

2.1 Data sources and search strategy

We conducted research on major bibliographic databases and publisher sites such as PubMed/Medline, IEEE Xplore, Scopus, and top journals related to image studies (e.g., Radiology, Medical Image Analysis, Lancet Digital Health, Nature Medicine) using articles that were roughly published between 2012 and early 2025. We selected this range because it spans when deep learning emerged as a technique for computer vision and was adapted into medical image analysis.

Search queries: “Deep learning” and associated concepts with architecture names (e.g., “deep learning,” “convolutional neural network,” “U-Net,” “transformer,” “self-supervised learning,” “foundation model”) with “CT,” “MRI,” “ultrasound,” “digital pathology,” “ophthalmology,” “radiology,” “prognostic modelling,” “reconstruction.” We additionally searched the citations of prominent review articles and top primary studies to find other relevant studies [1] – [3], [9] – [12].

2.2 Inclusion and Exclusion Criteria

study prioritized:

- Peer-reviewed articles with findings based on medical image data.
- Studies using deep learning techniques (such as CNNs, encoder-decoder networks, generative networks, transformers, or self-supervised/foundation models).

Contains:

- Application areas: major image-based technologies (radiography, CT, MRI, US, DP, OP image analysis, PET and hybrids).
- Data analysis includes classification, detection, segmentation, reconstruction, and risk models for prognosis.
- Articles with quantitative analysis related to clinically important outcomes or well-defined sets of comparison data.

study excluded:

- Studies only on non-medical image datasets.
- Purely methodological papers with no evaluation on medical imaging data.
- Reports with poor description of methodologies or unquantifiable target performances.

2.3 Limitations of the approach

Since this is a narrative review, it should not be expected that every study which relates to this systematic analysis will or should be included. Our own selection will necessarily be determined by search keywords, coverage of databases searched, and our own judgment about how important studies are—whether these be foundational studies on research methods or important clinical trials or other kinds of survey studies. There will not be a PRISMA flow diagram included with this article nor will there be any meta-analysis statistics provided.

3. Foundations of Deep Learning for Medical Imaging

3.1 From Classical Image Analysis to Deep Learning

Prior to deep learning, CAD systems were based on predefined features: grey level cooccurrence matrices, wavelet transforms, gradient-based features, shape features, and specialized features (plaque burden and tumour heterogeneity). Radiomics fully characterized high throughput extraction of hundreds or thousands of image features which could then be input into traditional machine learning models for classification or regression purposes [1], [2]. While these workflows showed that image contains rich quantitative features related to disease phenotype, these traditional models were plagued with challenges of handcrafted features that were sensitive to image acquisition variations and risk of overfitting when dealing with large numbers of features [1]–[3].

International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

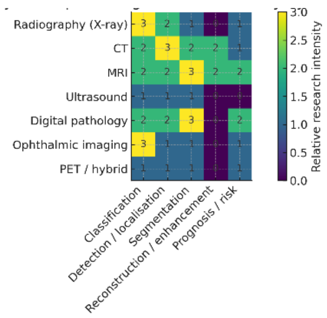

figure1. Taxonomy of deep learning for medical image analysis

Figure 1. Conceptual taxonomy of deep learning applications in medical image analysis, arranged by imaging modality (rows) and clinical task (columns). Dark shading represents combinations that have been most heavily studied in literature, such as chest radiography for disease classification, CT for lung nodule detection, MRI for brain tumour segmentation, and fundus photography for diabetic retinopathy screening. Lighter shading denotes areas where the evidence base is relatively sparse. This dual organisation mirrors the structure of the review and highlights both mature and under-explored regions of the current landscape.

Table 1. Representative public datasets

Dataset	Modality / domain	Primary task(s)	Approx. sample size	Key characteristics
ChestX-ray14	Chest radiography	Multi-label disease classification	112,120 frontal X-rays	14 text-mined labels; weakly supervised labels; early CXR benchmark.
CheXpert	Chest radiography	Multi-label disease classification	224,316 radiographs	Uncertainty-aware labels; strong baselines; widely used for CXR AI.
MIMIC-CXR	Chest radiography + text	Classification, report generation	377,110 images with free-text reports	Large de-identified CXR-report corpus; basis for vision–language and foundation models.
LUNA16	Chest CT	Pulmonary nodule detection	888 low-dose CT scans	Subset of LIDC-IDRI; benchmark for nodule detection.
BraTS	Brain MRI	Brain tumour segmentation	≈250–300 cases per year	Multi-modal MRI; standard benchmark for glioma segmentation.
ACDC	Cardiac MRI	LV/RV/myocardium segmentation; diagnosis	100 patients	Short-axis cine CMR; automated functional quantification.

Received: July 20, 2025

International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

fastMRI (knee)	Knee MRI (raw k-space)	Accelerated reconstruction	1,500+ multi-coil scans	Large public raw k-space dataset; deep reconstruction methods.
fastMRI (brain)	Brain MRI	Accelerated reconstruction	6,000+ scans	Multi-contrast brain MRI; reconstruction across field strengths.
EyePACS	Fundus photography	Diabetic retinopathy grading	80,000+ images	Large DR screening dataset; strong class imbalance.
CAMELYON16	Whole-slide histopathology	Metastatic breast cancer detection	400 WSIs	Multi-centre lymph node dataset; benchmark for pathology AI.

In medical imaging, early adopters have finetuned general computer vision CNNs-e.g., AlexNet, VGG, ResNet-on medical datasets and demonstrated significant improvements compared to classical radiomics for lesion classification and disease detection tasks [3], [9], [10].

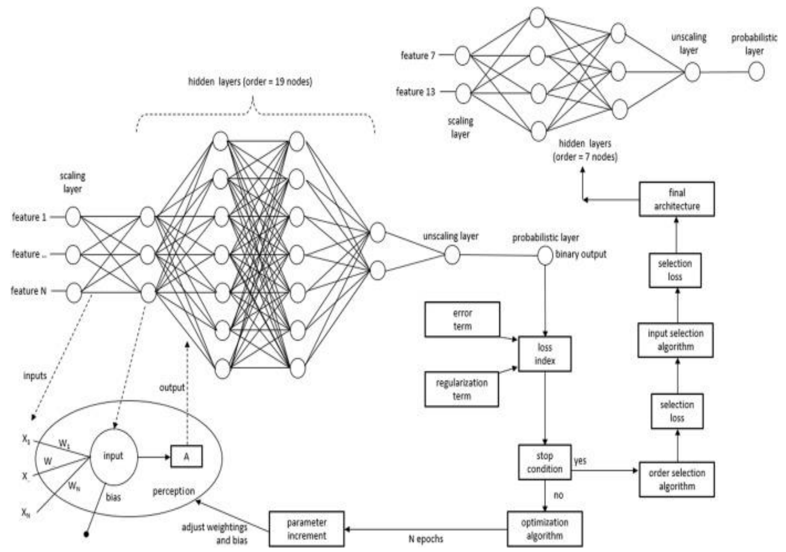

Figure 2 Overview of the anatomy of an ANN. ANN, artificial neural network [3].

Figure 2 An epoch is a single forward propagation (and back propagation through the ANN) of the entire data set. In small data sets this equals an iteration. In large data sets, the entire data set may not be able to be passed through the ANN in one batch. In this situation, the larger data set is broken into smaller batches. As each batch undergoes forward and back propagation, it is referred to as an iteration. Once all batches have been forward- and back-propagated this is an epoch[3]. over time, architecture specializes in handling 3D data, multi-channel inputs, and task-specific constraints such as shape priors and anatomical context.

International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.2 Core Deep Learning Architectures

Convolutional neural networks (2D and 3D): 2D and 3D. Standard 2D CNNs, inspired by the AlexNet and VGG models, are common in planar imaging. Extending the traditional 2D architecture to three dimensions to analyze volumetric data results in 3D CNNs. These, however, are more computationally intensive. Residual networks and DenseNets mitigate the problem of vanishing gradients by allowing for deeper architectures. Table 2 Key deep learning architecture families and their applications in medical image analysis.

Table 2. Core deep learning architectures

Architecture family	Key characteristics	Typical medical imaging uses	Example references / models
2D CNNs	Stacked convolution, pooling, non-linear layers	Image-level classification (e.g., DR, skin, chest radiography)	AlexNet, VGG, ResNet; CheXNet; DR screening networks
3D CNNs	3D convolutions capturing volumetric context	CT/MRI lesion detection and classification	3D lung nodule detectors; volumetric brain lesion classifiers
U-Net / encoder–decoder FCNs	Symmetric encoder–decoder with skip connections	Organ/tumour segmentation in CT, MRI, ultrasound	U-Net; nnU-Net family
GANs and related generative nets	Adversarial training with generator and discriminator	Low-dose CT denoising; cross-modality synthesis; data augmentation	GAN-based low-dose CT denoising; style-transfer for harmonisation
Transformers / ViT	Self-attention over image patches or tokens	Classification, segmentation, report generation	Vision Transformer variants; hybrid CNN–transformer models
Self-supervised models	Pretext tasks or contrastive objectives without labels	Pre-training for CT/MRI segmentation and classification	Models Genesis; SimCLR-style pre-training on medical volumes
Multi-modal foundation models	Large-scale pre-training on image–text pairs across modalities and languages	Zero-/few-shot classification, report generation, triage	Multi-modal, multi-lingual foundation models (e.g., M3FM-like)

International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Fully convolutional and encoder–decoder networks: Semantic segmentation tasks require predictions at a pixel- or voxel-level. FCNs involve convolutional layers for generating dense predictions. Quite logically, U-Net [7] introduced a symmetric structure of encoder-decoder with skip connections that became standard for biomedical image segmentation. nnU-Net showed that a well-designed self-configuring U-Net can achieve top performance on a wide variety of segmentation benchmarks without manual adaptation.[23]

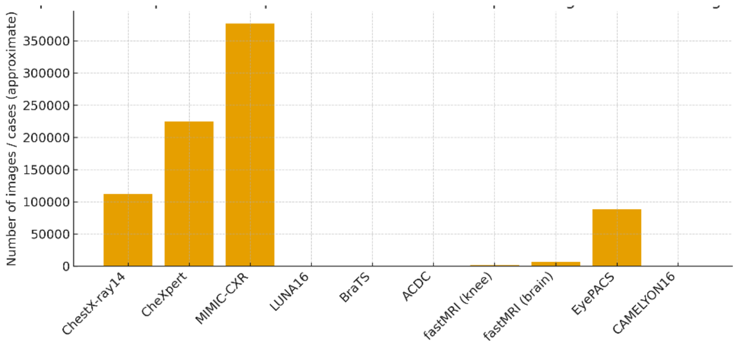

Figure 3. sample sizes of representative public datasets

Generative models: GANs, introduced by Goodfellow et al. [8], pit a generator network against a discriminator in a minimax game, enabling realistic image synthesis. In medical imaging, GANs have been applied to data augmentation, cross-modality synthesis (e.g., CT↔MRI), artefact reduction and reconstruction. For example, Yang et al. showed low-dose CT denoising using a GAN trained with perceptual and Wasserstein losses to preserve diagnostic detail while reducing noise [16]. Such generative approaches bring about both opportunities-e.g., dose reduction, harmonization-and risks-e.g., hallucination of structures-and therefore require rigorous task-based evaluations.

3.3 Emerging Paradigms

Transformers and vision transformers: Transformers were initially designed for natural language processing and are based on self-attention mechanisms to model long-range dependencies. Dosovitskiy et al. introduced the Vision Transformer, which segregates an image into patches to be processed by transformer layers [14]. Recently, there was increasing interest in using vision transformers or hybrid CNN-transformer models for medical imaging applications, such as classification, segmentation, and report generation, especially when pre-training on large-scale datasets can be done, for instance, on chest X-ray images or multimodal radiology-text datasets.

Self-supervised and semi-supervised learning: The problem of label scarcity represents a significant bottleneck within medical imaging. Self-supervised learning mitigates this by using pre-text tasks, requiring only raw images, such as contrastive learning, e.g., SimCLR [13], or restoration tasks. Models: Genesis proposed a suite of 3D inpainting, local shuffling, and other transformations to learn general-purpose 3D

International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

medical representations from unlabelled volumes, which can be fine-tuned for downstream tasks [15]. Semi-supervised learning generally combines small, labelled datasets with larger unlabelled corpora, often via consistency regularization or pseudo-labelling, and has shown particular promise in segmentation and detection in low-resource clinical settings.

3.4 Radiography and Chest Imaging (X-ray)

Transformers were originally developed for natural language processing and rely on self-attention mechanisms to model long-range dependencies. Dosovitskiy et al. proposed the Vision Transformer, which partitioned an image into patches to be processed via transformer layers [14]. There has been increasing exploration of pure vision transformers and hybrid CNN-transformer architectures in medical imaging for the purposes of classification, segmentation, and report generation, especially when large-scale pre-training is possible, for example on chest X-ray or multimodal radiology-text corpora.

Table 3. Modality-specific applications

Modality	Representative applications	Typical architectures	Main strengths	Key challenges
Radiography (X-ray)	CXR disease classification, triage, device detection	2D CNNs, transformers	Very large public datasets; high-throughput screening tasks	Dataset shift; underdiagnosis bias; limited localisation in pure classifiers
CT	Lung nodule detection; stroke/PE triage; low-dose denoising	3D CNNs, detection nets, GANs	Rich volumetric information; strong triage use cases	Multi-vendor heterogeneity; risk of hallucinated findings in reconstruction
MRI	Brain and cardiac segmentation; tumour characterisation; reconstruction	2D/3D U-Net, nnU-Net, unrolled reconstruction nets	Excellent soft-tissue contrast; strong benchmarks (BRATS, ACDC, fastMRI)	Protocol variability; translating challenge performance into clinical impact
Ultrasound	Echocardiographic view classification; foetal biometry; lesion detection	2D/3D CNNs, lightweight nets	Real-time imaging; potential for on-device AI	Operator dependence; variable image quality; limited multi-centre datasets

Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Digital pathology	Cancer detection and grading; mutation prediction	Patch-based CNNs, MIL, transformers	High-resolution morphology ; strong retrospective evidence	Heterogeneous staining and scanners; large WSI sizes; few prospective trials
Ophthalmic imaging	DR/AMD/glaucoma detection from fundus and OCT	2D CNNs, hybrid pipelines	Regulatory-approved AI examples; screening at scale	Maintaining performance in real-world programmes; fairness across populations
PET / hybrid	Lesion detection, staging, response prediction	3D CNNs, multi-modal fusion nets	Combined metabolic and anatomical information	Smaller datasets; tracer and scanner variability; harmonisation needs

However, Zech et al. demonstrated that the performance of a pneumonia detection network trained on an institution was significantly degraded when tested in external hospitals, pointing out critical issues related to dataset shift and spurious correlates (e.g., hospital-specific markers) [19]. Recent works focus on self-supervised and foundation models for chest X-ray, which can be fine-tuned across tasks, including abnormality detection, report generation, and survival prediction, and combined with clinical variables. The goal of these models is to offer a universal feature extractor for chest imaging in a robust way, but their clinical validation remains limited.

3.5 Computed Tomography (CT)

CT offers volumetric information at high spatial resolution and is often at the heart of diagnosis and management in oncology, lung disease, and acute care. The LUNA16 challenge standardized the evaluation of pulmonary nodule detection using low-dose chest CT. As a result, deep CNN-based CAD systems have outperformed earlier handcrafted methods [21]. Gruetzemacher et al. employed 3D CNNs to detect nodules and differentiate clinically significant lesions, demonstrating the advantage of volumetric context.[22]

3.6 Magnetic Resonance Imaging (MRI)

MRI offers excellent soft tissue contrast but suffers from long acquisition times and complex protocols. Deep learning has been widely applied to segmentation, quantification, and reconstruction. Havaei et al. proposed one of the early deep-learning pipelines using CNNs for brain tumour segmentation and demonstrated remarkable performance on the BRATS dataset [25]. The nnU-Net framework [23] by Isensee et

International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

al. achieved state-ofthe-art segmentation performance across many MRI segmentation challenges (brain, prostate, pancreas) through automatic adaptation of architecture and training hyperparameters to the dataset, with further emphasis on robust, taskagnostic baselines.

3.7 Ultrasound

Ultrasound is real-time, operator-dependent, and relatively inexpensive, making it both attractive and challenging for AI due to variable acquisition and artefacts. Madani et al. applied deep CNNs for classifying standard echocardiographic views (e.g., apical four-chamber, parasternal long-axis) from 2D echo cine loops, achieving accuracy comparable with cardiologists and reducing dependence on expert sonographers for view labelling [32]. Other work extends to foetal biometry, breast and thyroid nodule classification, vascular imaging, and point-of-care ultrasound, with deep learning applied to automate quality control, standardise measurements, and detect pathology.

Whole-slide images in pathology are extremely high-resolution and rich in morphological detail. Deep learning has shown remarkable success in the detection, grading, and mutation prediction of cancers. Coudray et al. showed that CNNs can classify non–small cell lung cancer subtypes and predict common driver mutations directly from H&E-stained WSIs [30]. Campanella et al. proposed weakly supervised multiple-instance learning to train slide-level cancer detection models using only slide-level labels (no pixel-level annotations), achieving clinical-grade performance across prostate, basal cell, and breast cancer cohorts.[31]

Image-level classification remains the most common application of deep learning in medical imaging. Various multi-label disease classification models such as CheXNet [17], CheXpert [18] have been trained for chest radiography on hundreds of thousands of images. In ophthalmology, DR detection algorithms [26, 27] and OCT-based referral systems [28] demonstrate high AUROC values and scalability for screening. In dermatology, Esteva et al. showed that a deep CNN trained on 129,450 clinical images could classify skin lesions at dermatologist-level accuracy.[29]

Segmentation underpins quantitative imaging and enables volumetry, shape analysis and radiomics. U-Net [7] and its variants, along with nnU-Net [23], dominate organ and lesion segmentation across modalities. For example, variations of U-Net have seen widespread use in the segmentation of lung, liver, and brain tumors, while nnU-Net has topped rankings in many public challenges due to its ability to automatically adapt to dataset characteristics [23], [25]. In chest radiography, one recent example is the work of Abedalla et al. who utilized U-Net with modern encoders (Xception, ResNet) to segment lung fields and pneumothorax regions, illustrating how encoder choice and pre-training can affect segmentation accuracy [45]. Structure-aware networks further integrate anatomical priors or shape regularisation to produce anatomically plausible segmentations, something of critical importance when segmenting complex structures (e.g., cardiac chambers, vessels).

International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 4. Tasks, metrics, and considerations

Task	Typical model outputs	Common evaluation metrics	Key clinical considerations
Classification (image-level)	Disease present/absent; multi-label findings	AUROC, AUPRC, accuracy, sensitivity, specificity	Threshold selection; calibration; impact on triage and workload; need for localisation
Detection / localisation	Bounding boxes or keypoints for lesions	Free-response ROC, sensitivity vs FP/image, mAP	False-positive burden; effect on reading time; interpretability of highlighted regions
Segmentation	Pixel/voxel-wise labels for structures/lesions	Dice coefficient, IoU, Hausdorff distance, surface distances	How segmentation errors affect volumetry, planning, and downstream decisions
Reconstruction / enhancement	Reconstructed images from undersampled or low-dose data	PSNR, SSIM, task-based image quality metrics	Preservation of subtle pathology; risk of hallucinated features; need for reader studies
Prognosis / risk stratification	Risk scores, predicted survival or response probabilities	C-index, calibration plots, time-dependent ROC	Handling censored data; integration with clinical covariates; decision-curve analysis; external validation

5. Clinical Translation, Validation, and Deployment

5.1 Datasets, Benchmarks, and External Validation

Publicly available datasets have formed the backbone for methodological advances and benchmarking, such as ChestX-ray14, CheXpert, MIMIC-CXR, LUNA16, BRATS, fastMRI [9], [18], [20], [21], [42]. However, performance on public test sets does not ensure real-world robustness. Zech et al. showed substantial performance drops when a pneumonia detection model was deployed across institutions with different patient populations and acquisition protocols [19]. These are key reasons why clinical translation requires:

- **External validation** across different sites, devices, and demographics.
- **Temporal validation** to account for evolving practice patterns and disease prevalence.

International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- **Application-specific metrics**, including calibration and decision-curve analysis, beyond AUROC.
- **Task-based reader studies** comparing AI-assisted versus unassisted clinicians.

Realistic assessment requires high-quality benchmarks, including multi-institutional data, standardized evaluation protocols, and clinically meaningful endpoints-all of which are still limited.

5.2 Robustness, Generalizability, and Fairness

Deep models are susceptible to domain shift (scanners, protocols, populations), label noise, and adversarial perturbations. Dataset shift has been described in chest radiography [19] and elsewhere, with models latching onto confounding cues (e.g. laterality markers, portable scanners). Fairness issues are increasingly recognized. Seyyed-Kalantari et al. showed underdiagnosis bias in chest X-ray models, with systematically worse performance in underserved groups (e.g. women, racial minorities) despite comparable overall metrics [34]. Paul et al. analyzed openly available CXR datasets and reported incomplete and inconsistent demographic reporting, precluding rigorous bias evaluation.[44]

5.3 Explainability and Trust in Medical AI

Explainability is often cited as essential for clinician trust, but the precise role of explainability is nuanced. Gradient-based saliency maps [38] and Grad-CAM [39] provide coarse localization of regions that are most influential for a decision and are widely used to generate heatmaps over medical images. However, those methods can be unstable and may highlight regions that are not causally related to the pathology.

More advanced methods include:

- Attention in transformers can be probed to understand which patches or tokens are emphasized.
- Concept-based explanations, which map predictions to clinically meaningful high-level concepts (e.g., “consolidation”, “mass”, “effusion”);
- Prototype and case-based explanations: here, models retrieve similar past cases to contextualize decisions.

Explainability will need to be assessed, not only technically, but also regarding its effect on clinician behavior in terms of over- or under-reliance on AI and diagnostic calibration. Human factors studies will be required to understand when explainability tools really enhance safety and performance.

6. Conclusion

Deep learning has undoubtedly marked a paradigm shift in medical image analysis. However, CNNs, U-Net–style architectures, generative models, and transformers have achieved expert-level performance in many well-defined diagnostic tasks across the modalities of radiography, CT, MRI, ultrasound, digital pathology, and ophthalmic imaging. These advances could grant earlier and more accurate diagnosis, improved workflow efficiency, and new quantitative biomarkers for prognosis and treatment planning.

At the same time, translation from benchmark performance to reliable clinical benefit is not automatic. Challenges around dataset shift, robustness, fairness, explainability, regulatory oversight, and workflow integration remain substantial. By contrast with prior surveys that focus predominantly on technical performance in specific domains, this review has sought to integrate foundational and emerging deep learning paradigms with a dual modality– and task-oriented perspective, and to foreground clinical translation, governance, and future directions. Addressing the remaining challenges requires rigorous methodology, multi-center collaboration, prospective evaluation, and a sustained focus on human–AI partnership. If these challenges are met, deep learning will not merely augment diagnostic medicine—it will help redefine it as a data-driven, continuously learning enterprise that delivers more precise, equitable, and timely care.

Funding: This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU252668]

References:

[1] R. J. Gillies, P. E. Kinahan, and H. Hricak, “Radiomics: Images are more than pictures, they are data,” *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[2] P. Lambin *et al.*, “Radiomics: the bridge between medical imaging and personalized medicine,” *Nat. Rev. Clin. Oncol.*, vol. 14, pp. 749–762, 2017.

[3] G. Currie, “Machine learning and deep learning in medical imaging: intelligent imaging,” *J. Med. Imaging Radiat. Sci.*, vol. 50, no. 4, pp. 477–487, 2019.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Adv. Neural Inf. Process. Syst.*, vol. 25, pp. 1097–1105, 2012.

[5] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[7] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.

[8] I. J. Goodfellow *et al.*, “Generative adversarial nets,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014, pp. 2672–2680.

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[9] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[10] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Z. Med. Phys.*, vol. 29, no. 2, pp. 102–127, 2019.

[11] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nat. Med.*, vol. 25, no. 1, pp. 44–56, 2019.

[12] C. J. Kelly *et al.*, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Med.*, vol. 17, no. 1, p. 195, 2019.

[13] T. Chen *et al.*, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[14] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[15] Z. Zhou *et al.*, "Models Genesis: Generic autodidactic models for 3D medical image analysis," *Med. Image Anal.*, vol. 67, p. 101840, 2021.

[16] Q. Yang *et al.*, "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss," *IEEE Trans. Med. Imaging*, vol. 37, no. 6, pp. 1348–1357, 2018.

[17] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," arXiv:1711.05225, 2017.

[18] J. Irvin *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI*, 2019.

[19] J. Zech *et al.*, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLoS Med.*, vol. 15, no. 11, p. e1002683, 2018.

[20] A. E. W. Johnson *et al.*, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data*, vol. 6, no. 1, p. 317, 2019.

[21] A. A. A. Setio *et al.*, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," *Med. Image Anal.*, vol. 42, pp. 1–13, 2017.

[22] R. Gruetzemacher, A. Gupta, and D. Paradice, "3D deep learning for detecting pulmonary nodules in CT scans," *J. Am. Med. Inform. Assoc.*, vol. 25, no. 10, pp. 1301–1310, 2018.

[23] F. Isensee *et al.*, "nnU-Net: A self-adapting framework for U-Net-based medical image segmentation," *Nat. Methods*, vol. 18, pp. 203–211, 2021.

[24] W. Bai *et al.*, "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks," *J. Cardiovasc. Magn. Reson.*, vol. 20, no. 1, p. 65, 2018.

[25] M. Havaei *et al.*, "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, 2017.

[26] V. Gulshan *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.

[27] D. S. W. Ting *et al.*, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *JAMA*, vol. 318, no. 22, pp. 2211–2223, 2017.

[28] J. De Fauw *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nat. Med.*, vol. 24, no. 9, pp. 1342–1350, 2018.

[29] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017.

[30] N. Coudray *et al.*, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nat. Med.*, vol. 24, pp. 1559–1567, 2018.

[31] E. Campanella *et al.*, "Clinical-grade computational pathology using weakly supervised deep learning on whole-slide images," *Nat. Med.*, vol. 25, pp. 1301–1309, 2019.

[32] A. Madani *et al.*, "Fast and accurate view classification of echocardiograms using deep learning," *NPJ Digit. Med.*, vol. 1, p. 6, 2018.

[33] F. Liu *et al.*, "A multimodal multidomain multilingual medical foundation model for zero shot clinical diagnosis," *NPJ Digit. Med.*, vol. 8, p. 86, 2025.

[34] A. Seyyed-Kalantari *et al.*, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nat. Med.*, vol. 27, pp. 2176–2182, 2021.

[35] X. Liu *et al.*, "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension," *BMJ*, vol. 370, p. m3164, 2020.

[36] S. Cruz Rivera *et al.*, "Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension," *Nat. Med.*, vol. 26, pp. 1351–1363, 2020.

[37] V. Sounderajah *et al.*, "The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence," *Nat. Med.*, 2025.

[38] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. ICLR Workshop*, 2014.

Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[39] R. R. Selvaraju *et al.*, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.

[40] M. S. Ghassemi, L. Oakden-Rayner, and A. L. Beam, “The medical algorithmic audit,” *Lancet Digit. Health*, vol. 3, no. 9, pp. e645–e648, 2021.

[41] K. Hammernik *et al.*, “Learning a variational network for reconstruction of accelerated MRI data,” *Magn. Reson. Med.*, vol. 79, no. 6, pp. 3055–3071, 2018.

[42] F. Knoll *et al.*, “fastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning,” *Radiol. Artif. Intell.*, vol. 2, no. 1, p. e190007, 2020.

[43] G. S. Hong *et al.*, “Overcoming the challenges in the development and implementation of artificial intelligence in radiology,” *Insights Imaging*, vol. 14, no. 1, p. 12, 2023.

[44] H. Y. Paul *et al.*, “Demographic reporting and representation in publicly available chest radiograph datasets,” *J. Am. Coll. Radiol.*, vol. 19, no. 4, pp. 506–513, 2022.

[45] A. Abedalla *et al.*, “Chest X-ray pneumothorax segmentation using U-Net with Xception and ResNet encoders,” *Sci. Rep.*, vol. 11, p. 11536, 2021.

[46] K. Chang *et al.*, “Residual convolutional neural network for the determination of IDH status in low- and high-grade gliomas from MR imaging,” *Clin. Cancer Res.*, vol. 24, no. 5, pp. 1073–1081, 2018.

[47] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, 2015.