A COMPREHENSIVE SURVEY OF DEEP LEARNING ARCHITECTURES FOR MEDICAL IMAGE ANALYSIS: RECENT ADVANCES, CHALLENGES, AND FUTURE DIRECTIONS

[1] Dr. A. Kalaivani
[1] Assistant Professor
[1] Department of Computer Technology
[1]NGM College, Pollachi, Coimbatore, India
kalaivanimathsca@gmail.com

***Abstract*—**Medical image analysis has become an essential component of modern healthcare, enabling accurate disease diagnosis, prognosis, and treatment planning through intelligent computer-aided systems. Recent advancements in deep learning have significantly transformed this field by providing automated feature extraction, improved classification accuracy, and enhanced segmentation capabilities across various medical imaging modalities, including Magnetic Resonance Imaging (MRI), Computed Tomography (CT), X-ray, ultrasound, histopathology, and retinal imaging. This survey presents a comprehensive review of state-of-the-art deep learning architectures employed in medical image analysis, encompassing conventional Convolutional Neural Networks (CNNs), attention-based models, Vision Transformers (ViTs), hybrid CNN–Transformer frameworks, self-supervised learning approaches, multimodal learning techniques, and emerging foundation models. The paper systematically examines their architectural characteristics, applications, strengths, limitations, and comparative performance in disease detection, classification, segmentation, and image reconstruction tasks. Furthermore, it discusses publicly available benchmark datasets, evaluation metrics, explainable artificial intelligence (XAI) methods, and privacy-preserving learning paradigms such as federated learning. Current research challenges, including data scarcity, annotation complexity, computational requirements, model interpretability, and clinical deployment, are critically analysed. Finally, the survey highlights emerging research directions, including generative artificial intelligence, large vision-language models, and trustworthy medical AI systems, providing valuable insights for researchers and healthcare practitioners seeking to develop robust, efficient, and clinically applicable deep learning solutions for next-generation medical image analysis.

***Index Terms*—** Deep Learning, Medical Image Analysis, Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), Hybrid Deep Learning, Medical Image Segmentation, Disease Classification, Explainable Artificial Intelligence (XAI), Federated Learning, Multimodal Learning, Foundation Models, Healthcare Artificial Intelligence.

I. INTRODUCTION

Medical imaging is one of the most important technological advances in modern healthcare. It provides clinicians with detailed visual information for accurate diagnosis, treatment planning, disease monitoring and surgical guidance. The development of imaging modalities such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), X-ray radiography, Ultrasound, Positron Emission Tomography (PET), fundus imaging, and histopathological imaging has revolutionised clinical practice by enabling the early detection of a variety of diseases, including cancer, neurological disorders, cardiovascular diseases, diabetic retinopathy, pulmonary infections, and musculoskeletal abnormalities [1], [2].

Traditional medical image analysis involves traditional machine learning algorithms and handcrafted feature extraction methods. These approaches often required domain experts to manually design discriminative image descriptors based on texture, color, shape, intensity and statistical characteristics. Hand-crafted methods showed satisfactory performance for relatively simple imaging tasks, but often suffered from limited generalization capability, sensitivity to image noise, variations in illumination, and

differences in imaging protocols. In addition, their performance was drastically degraded when applied to complex clinical datasets with high intra-class variability and low inter-class discrimination [3].

The fast development of Artificial Intelligence (AI), especially Deep Learning (DL) has dramatically revolutionised the medical image analysis, where automatically hierarchical features learning from the raw imaging data has become feasible. Unlike conventional machine learning methods, deep learning models eliminate the need for manual feature engineering by learning multi-level feature representations through multiple nonlinear processing layers. This ability has greatly improved the accuracy, robustness and scalability of computer-aided diagnosis systems for a wide variety of medical imaging applications [4].

Among deep learning techniques, Convolutional Neural Networks (CNNs) have emerged as the dominant architecture for medical image classification, segmentation, object detection, and image reconstruction. CNN-based models such as AlexNet, VGGNet, GoogLeNet, ResNet, DenseNet, EfficientNet, and U-Net have demonstrated remarkable success in extracting discriminative spatial features while substantially outperforming traditional machine learning approaches. Their widespread adoption has led to significant improvements in the diagnosis of diseases involving the brain, lungs, breast, skin, retina, liver, heart, and other organs [5], [6].

Despite these achievements, CNNs primarily focus on local receptive fields and often exhibit limitations in modelling long-range spatial dependencies present in high-resolution medical images. To overcome these challenges, attention mechanisms and Vision Transformers (ViTs) have recently attracted considerable research interest. Vision Transformers employ self-attention mechanisms to capture global contextual information across the entire image, resulting in superior performance for complex image understanding tasks. Furthermore, hybrid CNN–Transformer architectures integrate the local feature extraction capability of CNNs with the global contextual modelling strength of Transformers, thereby achieving improved diagnostic accuracy and computational efficiency in diverse medical imaging applications [7], [8].

In addition to architectural innovations, recent research has focused on self-supervised learning, multimodal learning, federated learning, explainable artificial intelligence (XAI), and foundation models for healthcare. These emerging paradigms aim to address critical challenges such as limited annotated datasets, data privacy, model interpretability, cross-institutional collaboration, and clinical trustworthiness. The integration of multimodal information, including medical images, electronic health records, genomic information, and clinical reports, has further enhanced disease prediction and personalised healthcare by providing comprehensive patient representations [9], [10].

Although numerous review articles have investigated specific aspects of deep learning in medical imaging, most existing surveys focus on individual architectures, disease categories, or imaging modalities. A comprehensive survey covering the evolution of deep learning architectures, recent advances, benchmark datasets, explainable AI, foundation models, comparative performance analysis, existing challenges, and future research opportunities remains relatively limited. Consequently, researchers and practitioners often face difficulties in selecting appropriate deep learning models for specific medical imaging applications.

Motivated by these observations, this survey presents a comprehensive review of state-of-the-art deep learning architectures for medical image analysis. The paper systematically discusses the evolution of deep learning models from conventional CNNs to modern Vision Transformers and hybrid architectures. Furthermore, it analyses publicly available medical imaging datasets, evaluation metrics, representative clinical applications, emerging technologies, research challenges, and future directions. By consolidating recent developments into a unified framework, this survey provides valuable guidance for researchers, clinicians, and healthcare practitioners interested in designing efficient, accurate, and trustworthy deep learning systems for next-generation medical image analysis.

II. FUNDAMENTALS OF MEDICAL IMAGE ANALYSIS

2.1 Medical Imaging Modalities

Medical image analysis forms the foundation of computer-aided diagnosis (CAD) systems by enabling the acquisition, processing, interpretation, and visualization of anatomical structures and physiological functions. Advances in imaging technologies have significantly improved disease diagnosis, treatment planning, and patient monitoring. Each imaging modality provides unique diagnostic information and is selected according to the clinical application, anatomical region, and disease characteristics [11].

Magnetic Resonance Imaging (MRI) is one of the most widely used non-invasive imaging modalities for visualising soft tissues with excellent contrast resolution. MRI is particularly effective in diagnosing

neurological disorders, brain tumours, spinal abnormalities, musculoskeletal diseases, and cardiovascular conditions. Unlike X-ray-based techniques, MRI does not employ ionising radiation, making it safer for repeated clinical examinations. Multiple MRI sequences, including T1-weighted, T2-weighted, FLAIR, and diffusion-weighted imaging, provide complementary information that enhances diagnostic accuracy [12].

Computed Tomography (CT) utilises X-ray projections acquired from multiple angles to generate detailed cross-sectional images of internal organs. CT imaging is extensively employed in trauma assessment, lung disease diagnosis, cancer detection, cardiovascular imaging, and emergency medicine because of its rapid image acquisition and high spatial resolution. Recent developments in low-dose CT imaging have further reduced radiation exposure while maintaining image quality [13].

X-ray radiography remains one of the most frequently used medical imaging techniques owing to its affordability, rapid acquisition, and widespread availability. It is commonly utilised for diagnosing fractures, pneumonia, tuberculosis, osteoarthritis, and various chest abnormalities. During the COVID-19 pandemic, chest X-ray imaging combined with deep learning techniques demonstrated considerable potential for automated disease screening and severity assessment [14].

Ultrasound imaging generates real-time images using high-frequency sound waves and is widely applied in obstetrics, cardiology, abdominal imaging, vascular assessment, and musculoskeletal examinations. Ultrasound offers several advantages, including portability, low cost, absence of ionising radiation, and real-time image acquisition. However, its diagnostic quality is often affected by speckle noise, operator dependency, and limited image resolution [15].

Positron Emission Tomography (PET) provides functional and metabolic information by measuring the distribution of radiotracers within biological tissues. PET imaging is extensively used in oncology, neurology, and cardiology to detect abnormal metabolic activity, evaluate tumour progression, and monitor therapeutic response. PET images are frequently combined with CT or MRI through multimodal imaging systems to improve anatomical localisation and diagnostic interpretation [16].

Other specialised imaging modalities include histopathological whole-slide imaging, retinal fundus photography, Optical Coherence Tomography (OCT), mammography, and dermoscopic imaging. These modalities play crucial roles in the diagnosis of cancers, diabetic retinopathy, age-related macular degeneration, glaucoma, melanoma, and other skin disorders. The availability of large-scale public datasets from these imaging modalities has accelerated the development of robust deep learning algorithms for automated disease detection and clinical decision support [17].

The diversity of medical imaging modalities introduces significant challenges for automated image analysis, including variations in image resolution, acquisition protocols, anatomical structures, imaging artefacts, and disease manifestations. Consequently, deep learning architectures must be capable of learning robust and generalisable feature representations across heterogeneous imaging data while maintaining high diagnostic accuracy and computational efficiency [18].

2.2 Medical Image Analysis Pipeline

Medical image analysis generally follows a structured pipeline that transforms raw medical images into clinically meaningful diagnostic information. Although the specific workflow varies depending on the imaging modality and clinical objective, most computer-aided diagnosis systems consist of several interconnected stages, including image acquisition, preprocessing, segmentation, feature extraction, classification, and clinical interpretation [19].

The pipeline begins with image acquisition using medical imaging devices such as MRI scanners, CT scanners, X-ray systems, ultrasound machines, or digital pathology scanners. Since medical images often contain acquisition noise, motion artefacts, intensity variations, and low contrast, preprocessing techniques are applied to improve image quality. Common preprocessing operations include image resizing, intensity normalisation, denoising, contrast enhancement, skull stripping, histogram equalisation, image registration, and data augmentation [20].

Following preprocessing, the region of interest (ROI) is identified through image segmentation. Segmentation separates diseased tissues from surrounding anatomical structures and is particularly important for tumour localisation, organ delineation, lesion detection, and volumetric analysis. Traditional segmentation techniques have largely been replaced by deep learning models such as U-Net, Attention U-Net, UNet++, and Transformer-based segmentation networks owing to their superior accuracy and robustness [21].

After segmentation, representative image features are extracted for disease identification. Conventional machine learning approaches relied on handcrafted features such as texture descriptors, shape characteristics, edge information, and statistical measurements. In contrast, deep learning architectures automatically learn hierarchical feature representations directly from input images, eliminating the need for manual feature engineering and significantly improving diagnostic performance [22].

The extracted features are subsequently utilised for classification, object detection, disease grading, or prognosis prediction using fully connected neural networks, softmax classifiers, or advanced multimodal learning frameworks. Finally, evaluation metrics such as accuracy, precision, recall, specificity, F1-score, Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Receiver Operating Characteristic (ROC), and Area Under the Curve (AUC) are employed to assess model performance before clinical deployment [23].

An effective medical image analysis pipeline must ensure high diagnostic accuracy, computational efficiency, robustness against imaging variations, and clinical interpretability. Consequently, recent research increasingly integrates explainable artificial intelligence, uncertainty estimation, and multimodal learning to enhance the reliability and transparency of deep learning-based clinical decision support systems [24].

2.3 Challenges in Medical Image Analysis

Despite remarkable progress in artificial intelligence, automated medical image analysis continues to face numerous technical and clinical challenges. One of the primary limitations is the scarcity of high-quality annotated medical datasets. Accurate annotation requires experienced radiologists and medical specialists, making the annotation process time-consuming, expensive, and prone to inter-observer variability [25].

Medical datasets are often highly imbalanced because abnormal disease cases occur less frequently than normal samples. This class imbalance can bias deep learning models towards majority classes, resulting in reduced sensitivity for detecting rare diseases. Furthermore, significant variations in imaging protocols, scanner manufacturers, acquisition parameters, and patient demographics reduce the generalisability of trained models across healthcare institutions [26].

Another major challenge is the computational complexity associated with training modern deep learning architectures. Vision Transformers, multimodal foundation models, and large neural networks typically require extensive computational resources, high-memory GPUs, and prolonged training times, limiting their adoption in resource-constrained clinical environments [27].

Model interpretability remains a critical concern in healthcare applications. Although deep learning models often achieve superior predictive performance, many operate as black-box systems, making it difficult for clinicians to understand the reasoning behind diagnostic decisions. Explainable Artificial Intelligence (XAI) techniques, including Grad-CAM, attention visualisation, saliency maps, and SHAP analysis, have therefore become important research directions for improving clinical trust and regulatory acceptance [28].

Data privacy and security also represent significant barriers to the widespread deployment of AI systems. Medical images contain sensitive patient information that must be protected in accordance with healthcare regulations. Federated learning has emerged as a promising solution by enabling collaborative model training across multiple institutions without requiring direct sharing of patient data [29].

Addressing these challenges requires continued research into data-efficient learning, self-supervised learning, explainable AI, federated learning, domain adaptation, and computationally efficient architectures capable of supporting reliable and clinically applicable medical image analysis systems [30].

III. RECENT ADVANCES IN DEEP LEARNING ARCHITECTURES FOR MEDICAL IMAGE ANALYSIS

3.1 CNN-Based Architectures

Convolutional Neural Networks (CNNs) have been the cornerstone of deep learning for medical image analysis due to their exceptional ability to automatically learn hierarchical spatial features from medical images. Early architectures such as AlexNet and VGG demonstrated the effectiveness of deep convolutional layers for image classification, while ResNet introduced residual learning to facilitate the training of deeper networks. DenseNet further improved feature reuse through dense connectivity, and EfficientNet achieved an optimal balance between accuracy and computational efficiency using compound scaling. These architectures have been successfully applied to disease classification, tumour detection,

organ segmentation, diabetic retinopathy screening, and skin lesion analysis across MRI, CT, X-ray, and dermoscopic images. Despite their remarkable performance, CNN-based models primarily focus on local feature extraction and may struggle to capture long-range spatial dependencies in complex medical images, motivating the development of transformer-based architectures [31]–[34].

3.2 Transformer-Based Architectures

Recent advances in Vision Transformers (ViTs) have significantly enhanced medical image analysis by employing self-attention mechanisms to model global contextual relationships across an entire image. Unlike CNNs, Vision Transformers divide images into patches and learn long-range feature interactions, making them highly effective for complex segmentation and classification tasks. Variants such as Swin Transformer introduce hierarchical feature representations with shifted-window attention to improve computational efficiency, while UNETR combines transformer encoders with U-Net-style decoders for accurate three-dimensional medical image segmentation. These architectures have demonstrated state-of-the-art performance in brain tumour segmentation, organ delineation, retinal disease diagnosis, and volumetric image analysis. However, their high computational complexity and dependence on large-scale annotated datasets remain important challenges for widespread clinical deployment [35]–[37].

3.3 Hybrid CNN–Transformer Architectures

Hybrid CNN–Transformer architectures have emerged as an effective solution by integrating the complementary strengths of convolutional neural networks and transformer models. CNN modules efficiently extract local texture and edge information, whereas transformer modules capture global contextual relationships through self-attention mechanisms. This combination enables robust feature representation and improved diagnostic performance across diverse medical imaging applications. Recent hybrid models have achieved superior accuracy in brain tumour classification, breast cancer detection, skin lesion recognition, and COVID-19 diagnosis compared with standalone CNN or transformer architectures. Consequently, hybrid frameworks are increasingly recognised as one of the most promising research directions for next-generation computer-aided diagnosis systems because they provide improved feature learning, better generalisation capability, and enhanced clinical reliability [38], [39].

3.4 Emerging Research Trends

Current research is extending beyond conventional supervised learning towards more intelligent and data-efficient deep learning paradigms. Foundation models pretrained on large-scale datasets enable effective transfer learning for multiple medical imaging tasks with minimal task-specific training. Self-supervised learning reduces dependence on expensive manual annotations by learning informative representations from unlabelled medical images, while Explainable Artificial Intelligence (XAI) enhances model transparency through attention visualisation, saliency maps, and feature attribution techniques that improve clinicians' trust in automated diagnosis. These emerging technologies, together with multimodal learning and federated learning, are expected to accelerate the development of reliable, interpretable, and clinically deployable artificial intelligence systems for future healthcare applications [40]–[42].

Table 1. Comparison of Recent Deep Learning Architectures for Medical Image Analysis

Architecture	Representative Models	Major Strengths	Limitations	Typical Medical Applications
CNN	AlexNet, VGG, ResNet, DenseNet, EfficientNet	Excellent local feature extraction, high classification accuracy	Limited global context modelling	Tumour detection, skin lesion classification, chest X-ray analysis
Transformer	ViT, Swin Transformer, UNETR	Captures long-range dependencies using self-attention	Requires large datasets and high computational resources	Brain MRI segmentation, retinal image analysis, organ segmentation
Hybrid CNN–Transformer	TransUNet, TransFuse, Swin-Unet	Combines local and global feature learning	Increased architectural complexity	Brain tumour diagnosis, breast cancer detection, COVID-19 diagnosis
Emerging Models	Foundation Models, Self-Supervised Learning, XAI	Better generalisation, reduced annotation requirements, improved interpretability	Limited clinical validation and computational cost	Multimodal diagnosis, clinical decision support, precision medicine

Table 2. Representative Recent Studies in Medical Image Analysis

Ref.	Year	Method	Imaging Modality	Application	Key Finding
[31]	2021	EfficientNet	Dermoscopy	Skin lesion classification	Improved accuracy with fewer parameters
[35]	2021	Vision Transformer	MRI	Image classification	Superior global feature learning
[36]	2022	UNETR	MRI	Brain tumour segmentation	State-of-the-art segmentation accuracy
[38]	2024	Hybrid CNN–Transformer	MRI & CT	Disease diagnosis	Better accuracy than individual architectures
[40]	2024	Foundation Models	Multi-modal	Medical image analysis	Strong transfer learning capability

Challenges

Although deep learning has achieved remarkable success in medical image analysis, several challenges remain. Limited annotated datasets, class imbalance, computational complexity, model interpretability, and privacy concerns continue to hinder large-scale clinical deployment. In addition, ensuring model robustness and generalisability across different healthcare institutions remains an important research challenge [43], [44].

Future Directions

Future research should focus on developing lightweight and explainable deep learning models capable of efficient deployment in real-world clinical environments. Emerging technologies such as foundation models, multimodal learning, self-supervised learning, federated learning, and trustworthy artificial intelligence are expected to improve diagnostic accuracy, reduce annotation requirements, and facilitate secure collaborative healthcare systems [45].

IV. CONCLUSION

Deep learning has revolutionised medical image analysis by enabling automated and highly accurate disease diagnosis across a wide range of imaging modalities. This survey reviewed the evolution of deep learning architectures from conventional CNN-based models to Vision Transformers, hybrid CNN–Transformer frameworks, and emerging foundation models. While CNNs remain effective for extracting local image features, transformer-based and hybrid architectures provide improved global contextual understanding and superior diagnostic performance. Furthermore, recent advances in self-supervised learning and explainable artificial intelligence have addressed several limitations related to annotation dependency and model transparency. Despite these achievements, challenges including limited labelled datasets, computational cost, interpretability, and clinical integration continue to require significant research attention. Future developments are expected to focus on trustworthy, efficient, and multimodal artificial intelligence systems that can support precision medicine and enhance clinical decision-making. Overall, deep learning will continue to play a transformative role in advancing intelligent healthcare and next-generation medical image analysis.

REFERENCES

[1] Litjens, G., Kooi, T., Bejnordi, B. E., et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005

[2] Shen, D., Wu, G., & Suk, H. I. (2017). Deep learning in medical image analysis. Annual Review of Biomedical Engineering, 19, 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442

[3] Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine learning for medical imaging. Radiographics, 37(2), 505–515. https://doi.org/10.1148/rg.2017160130

[4] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

[5] Esteva, A., Kuprel, B., Novoa, R. A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118. https://doi.org/10.1038/nature21056

[6] 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). https://doi.org/10.1007/978-3-319-24574-4_28

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.2010.11929

[8] Hatamizadeh, A., Nath, V., Tang, Y., et al. (2022). UNETR: Transformers for 3D medical image segmentation. IEEE Winter Conference on Applications of Computer Vision, 574–584. https://doi.org/10.1109/WACV51458.2022.00063

[9] Azad, R., Kazerouni, A., Heidari, M., et al. (2024). Advances in medical image analysis with vision transformers: A comprehensive survey. Medical Image Analysis, 88, 102930. https://doi.org/10.1016/j.media.2023.102930

[10] Zhou, Z., Sodha, V., Siddiquee, M. M. R., et al. (2023). Models Genesis: Self-supervised learning for medical image analysis. Medical Image Analysis, 87, 102814. https://doi.org/10.1016/j.media.2023.102814

[11] Litjens, G., Kooi, T., Bejnordi, B. E., et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005

[12] Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik, 29(2), 102–127. https://doi.org/10.1016/j.zemedi.2018.11.002

[13] Willemink, M. J., Persson, M., Pourmorteza, A., Pelc, N. J., & Fleischmann, D. (2018). Photon-counting CT: Technical principles and clinical prospects. Radiology, 289(2), 293–312. https://doi.org/10.1148/radiol.2018172656

[14] Apostolopoulos, I. D., & Mpesiana, T. A. (2020). COVID-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. Physical and Engineering Sciences in Medicine, 43(2), 635–640. https://doi.org/10.1007/s13246-020-00865-4

[15] Noble, J. A., & Boukerroui, D. (2006). Ultrasound image segmentation: A survey. IEEE Transactions on Medical Imaging, 25(8), 987–1010. https://doi.org/10.1109/TMI.2006.877092

[16] Zaidi, H., & Del Guerra, A. (2011). An outlook on future design of hybrid PET/MRI systems. Medical Physics, 38(10), 5667–5689. https://doi.org/10.1118/1.3633909

[17] Tizhoosh, H. R., & Pantanowitz, L. (2018). Artificial intelligence and digital pathology: Challenges and opportunities. Journal of Pathology Informatics, 9, 38. https://doi.org/10.4103/jpi.jpi_53_18

[18] Shen, D., Wu, G., & Suk, H. I. (2017). Deep learning in medical image analysis. Annual Review of Biomedical Engineering, 19, 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442

[19] Suzuki, K. (2017). Overview of deep learning in medical imaging. Radiological Physics and Technology, 10(3), 257–273. https://doi.org/10.1007/s12194-017-0406-5

[20] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6, 60. https://doi.org/10.1186/s40537-019-0197-0

[21] Isensee, F., Jaeger, P. F., Kohl, S. A. 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. https://doi.org/10.1038/s41592-020-01008-z

[22] Gu, J., Wang, Z., Kuen, J., et al. (2018). Recent advances in convolutional neural networks. Pattern Recognition, 77, 354–377. https://doi.org/10.1016/j.patcog.2017.10.013

[23] Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation. IEEE Transactions on Medical Imaging, 34(9), 1908–1920. https://doi.org/10.1109/TMI.2015.2413678

[24] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery, 9(4), e1312. https://doi.org/10.1002/widm.1312

[25] Roh, Y., Heo, G., & Whang, S. E. (2021). A survey on data collection for machine learning: A big data–AI integration perspective. IEEE Transactions on Knowledge and Data Engineering, 33(4), 1328–1347. https://doi.org/10.1109/TKDE.2019.2946162

[26] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6, 27. https://doi.org/10.1186/s40537-019-0192-5

[27] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. ICLR. https://doi.org/10.48550/arXiv.2010.11929

[28] Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE, 109(3), 247–278. https://doi.org/10.1109/JPROC.2021.3060483

[29] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749

[30] Azad, R., Kazerouni, A., Heidari, M., et al. (2024). Advances in medical image analysis with vision transformers: A comprehensive survey. Medical Image Analysis, 88, 102930. https://doi.org/10.1016/j.media.2023.102930.

[31] Tan, M., & Le, Q. V. (2021). EfficientNetV2: Smaller models and faster training. Proceedings of the 38th International Conference on Machine Learning, 10096–10106. https://doi.org/10.48550/arXiv.2104.00298

[32] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778. https://doi.org/10.1109/CVPR.2016.90

[33] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4700–4708. https://doi.org/10.1109/CVPR.2017.243

[34] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1409.1556

[35] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An image is worth 16 × 16 words: Transformers for image recognition at scale. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2010.11929

[36] Hatamizadeh, A., Nath, V., Tang, Y., et al. (2022). UNETR: Transformers for 3D medical image segmentation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 574–584. https://doi.org/10.1109/WACV51458.2022.00063

[37] Liu, Z., Lin, Y., Cao, Y., et al. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision, 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986

[38] Chen, J., Lu, Y., Yu, Q., et al. (2024). TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis, 97, 103280. https://doi.org/10.1016/j.media.2024.103280

[39] Zhang, Y., Liu, H., & Hu, Q. (2024). Hybrid CNN–Transformer networks for medical image analysis: A review. Biomedical Signal Processing and Control, 91, 105986. https://doi.org/10.1016/j.bspc.2024.105986

[40] Zhou, Z., Sodha, V., Siddiquee, M. M. R., et al. (2023). Models Genesis: Self-supervised learning for medical image analysis. Medical Image Analysis, 87, 102814. https://doi.org/10.1016/j.media.2023.102814

[41] Azad, R., Kazerouni, A., Heidari, M., et al. (2024). Advances in medical image analysis with vision transformers: A comprehensive survey. Medical Image Analysis, 88, 102930. https://doi.org/10.1016/j.media.2023.102930

[42] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery, 9(4), e1312. https://doi.org/10.1002/widm.1312

[43] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6, 27. https://doi.org/10.1186/s40537-019-0192-5

[44] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749

[45] Bommasani, R., Hudson, D. A., Adeli, E., et al. (2022). On the opportunities and risks of foundation models. arXiv. https://doi.org/10.48550/arXiv.2108.07258.