communications engineering

A Nature Portfolio journal

Review article

https://doi.org/10.1038/s44172-026-00602-x

Bridging modalities with AI: a review of AI advances in multimodal biomedical imaging

Le Minh Thao Doan [1], Kaveh Shahhosseini [1], Suraj Verma [1], Abdolreza Marefat[2], Giorgio Locicero [3], Sneha Verma[1], Claudio Angione[1,4,5] & Annalisa Occhipinti [1,4,5]

The rapid evolution of AI has facilitated innovative solutions in analysing different biomedical imaging modalities. By leveraging the complementary information from each modality, multimodal AI solutions have shown a huge potential to go beyond human capabilities and offer advances in bioimaging. At the same time, new foundation models and transformer-based architectures are now poised to address unsolved challenges in this field. This review aims to explore and discuss the state-of-the-art AI techniques applied in multimodal biomedical imaging, presenting the key challenges and future directions. We discuss several integration strategies to combine multiple biomedical imaging data types. We also focus on methods to overcome the open challenges related to data quality, model interpretability, and ethical implications.

Imaging plays a crucial role in biomedical research as it provides valuable information about functional processes at the tissue, cellular, and molecular levels[1], and its applications are widely used in early disease detection and therapeutic response monitoring[2–4]. For instance, several non-invasive imaging techniques, including X-rays, Computed Tomography (CT) scans, Magnetic Resonance Imaging (MRI), and Optical Coherence Tomography (OCT), are regularly used for disease diagnosis, disease progression monitoring, and therapeutic efficacy assessment[2,3,5]. Beyond disease diagnostics and monitoring, high-resolution biomedical imaging (e.g., microscopic and spectroscopic imaging) has recently provided new opportunities to investigate several biological processes, including the inner workings of cells and the tumour microenvironment, which plays a critical role in disease progression and treatment monitoring.

Traditionally, individual imaging techniques have provided unique perspectives on biological tissues and disease states. However, considering a single modality may not provide a comprehensive view of the disease. Hence, the need to learn from different modalities has driven the shift towards multimodal biomedical imaging techniques, where different imaging techniques are integrated to provide a holistic view of the biological mechanism of the disease[1]. By integrating patient-specific, structural, functional, and molecular insights from various imaging types, multimodal biomedical imaging frameworks can leverage complementary information from each modality to enhance the understanding of biological processes and diagnostic accuracy[6]. This integrated approach has the potential to drive the next frontier in precision medicine, offering more robust tools for disease characterisation and paving the way for tailored therapeutic strategies.

Recent advancements in Artificial Intelligence (AI) have reshaped the landscape of biomedical imaging, offering solutions that bridge various imaging modalities to deliver more precise diagnostics and facilitate precision medicine[7,8]. Specifically, AI has been used widely to analyse biomedical images from segmenting regions of interest to diagnosis, prognosis, and response to treatment[4,9].

AI techniques, particularly deep learning models (e.g., Convolutional Neural Networks (CNNs), Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and transformers), have demonstrated an exceptional ability to automate complex image analysis tasks, extract intricate information, and infer meaningful conclusions[4,10,11]. AI has also proven capable of handling massive and high-resolution data, while enabling the integration of information from various imaging types[7]. In fact, by leveraging unique insights from each modality, the application of AI in a multimodal bioimaging context can enhance the model performance and provide a comprehensive view of diseases. Recently, foundation models have also been proposed to enable universal medical image tasks, enabling accurate predictions across multiple tasks, including disease diagnosis and tumour segmentation[8,12].

Due to the continuous evolution of AI techniques to integrate different biomedical imaging data types, the current literature lacks a comprehensive review of the latest AI integration approaches for multimodal biomedical

[1]School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, UK. [2]Department of Computer Engineering, Technical and Engineering Faculty, South Tehran Branch, Islamic Azad University, Tehran, Iran. [3]Department of Mathematics and Computer Science, University of Catania, Catania, Italy. [4]Centre for Digital Innovation, Teesside University, Middlesbrough, UK. [5]National Horizons Centre, Teesside University, Darlington, UK.
e-mail: a.occhipinti@tees.ac.uk

data. For instance, many surveys have mainly focused on specific aspects in biomedical applications, including model-specific innovations (e.g., foundation and vision-language foundation models[13,14]), imaging-specific modality (e.g., radiological imaging[15] and histopathological imaging[16]), and fusion-specific strategies (e.g., intermediate fusion[17]). In contrast, our review provides a comprehensive overview of several imaging fusion strategies (e.g., pixel, feature, decision, and hierarchical-level approaches) for integrating multiple types of biomedical imaging (e.g., radiological, optical, microscopic, and spectroscopic modalities). While other review studies focus on multimodal AI across both imaging and non-imaging modalities (e.g., electronic health records (EHRs) and omics)[18,19], our work focuses on integration strategies and technical issues specifically tailored to multimodal biomedical imaging. Beyond this, we provide a detailed discussion of how complementary structural, functional, and molecular imaging can be exploited to improve the clinical decision-making process and outline emerging directions in the field.

Specifically, this review focuses on multimodal biomedical imaging, highlighting AI applications in image-to-image integration. Additionally, we address and discuss technical challenges unique to integrating imaging data (e.g., modality heterogeneity, image enhancement, spatial misalignment, and data augmentation methods), which are often underexplored in general multimodal reviews. We then examine and discuss several integration strategies to merge multiple biomedical imaging types, highlighting key challenges and proposing potential solutions to improve data fusion, model performance, and clinical applicability. Moreover, ethical considerations with explainability in multimodal AI approaches are also discussed, offering a unique perspective on interpretability and ethical challenges. To contextualise these advances, we review and discuss the major AI developments over the past fifteen years, highlighting the AI innovation and emerging role of multimodal integration in biomedical imaging. Finally, we present the future directions, including multimodal large language models (MLLMs) and specialist AI agents, to further enhance the successful applications of AI in multimodal biomedical imaging. By integrating technical, clinical, and ethical perspectives, our work presents a holistic view of multimodal AI applications, addresses current challenges, and guides future research in biomedical studies, thereby offering new insights and directions for the field.

Biomedical imaging and the shift towards multimodal AI techniques

Multimodal biomedical imaging includes a variety of imaging techniques, each contributing specific advantages in capturing structural, functional, and molecular aspects of the disease under investigation (as detailed in Box 1). Each biomedical imaging modality has unique characteristics, and by combining them, multimodal biomedical imaging enhances our understanding of biological processes at the organ (patient), tissue, and cellular levels, while improving diagnostic accuracy.

At the patient level, radiological imaging (Fig. 1A), including X-rays, ultrasound, CT, and MRI, plays a critical role in healthcare by characterising structural and morphological information and functional activity within tissues and offering a foundation for comprehensive diagnostic and prognostic assessments[20]. While the use of single-modality radiological images remains crucial to investigate specific anatomical and functional characteristics, their integration has recently become necessary due to the growing complexity of clinical cases and the need for comprehensive diagnostic insights. For example, ultrasound, a non-invasive diagnostic technique that uses high-frequency sound waves to generate real-time images, is optimal for multimodal AI applications, where its immediate feedback is combined with other radiological images (e.g., CT or MRI) to improve interventions (e.g., guiding biopsy)[21]. Additionally, Positron Emission Tomography (PET) and Single-Photon Emission Computed Tomography (SPECT) are nuclear imaging techniques that provide information on metabolic and functional activities, such as glucose uptake, which is indicative of cellular metabolism and is often high in cancerous tissues. The integrated analysis of PET with MRI and CT scan images provides detailed metabolic and functional information, complementing the high-resolution anatomical detail offered by MRI and CT and allowing the localisation of lesions and identification of structural abnormalities[22]. Hence, the integration of these image modalities leverages the strengths of each imaging type to generate a holistic picture of both structural and functional information, improving diagnostic precision[23].

While radiological imaging provides anatomical information at the patient-level, microscopic (e.g., histological and histopathological) and spectroscopic imaging (e.g., Raman, fluorescence, infrared, hyperspectral imaging, and Mass Spectrometry Imaging (MSI)) offer high-resolution views on the morphology and molecular information at the tissue level[24,25].

Box 1 | Bioimaging Modalities

1. Radiological imaging

Radiological imaging is one of the most common biomedical imaging modalities used in healthcare for diagnosis and treatment monitoring. These techniques can characterise anatomical and functional features of both tissues and organs.

Ultrasound imaging is a non-invasive diagnostic technique based on high-frequency sound waves to create real-time images.

Magnetic Resonance Imaging (MRI) employs magnetic fields and radio waves to generate highly detailed soft-tissue images.

Computed Tomography (CT) uses X-rays to produce cross-sectional images of the body's structures (e.g., bones, organs, and blood vessels).

Positron Emission Tomography (PET) and Single-Photon Emission Computed Tomography (SPECT) are nuclear imaging techniques that provide information on metabolic and functional activity in the body.

2. Microscopic and spectroscopic imaging

Microscopic imaging provides detailed information on tissue morphology, while spectroscopic techniques offer insights into molecular-level and chemical compositions.

Histological imaging involves the microscopic study of tissue anatomy to understand the organisation and structure of normal tissues.

Histopathological imaging, including whole slide images, is used to study disease-related changes at the tissue level.

Raman spectroscopy is a non-invasive and label-free technique using inelastic scattering of light to produce a molecular fingerprint of a sample.

Fluorescence imaging employs fluorescent markers by illuminating the sample with a light source of a specific wavelength to highlight specific components (e.g., proteins, biomolecules, and nanoparticles) within cells and tissues.

3. Optical imaging

Optical imaging modalities utilise light to detect anomalies within tissues, providing insights into tissue structure, physiology, and biochemical processes.

Optical coherence tomography (OCT) is an optical method that uses light waves to capture high-resolution cross-sectional images of biological tissues. OCT can capture intricate structural details, allowing the detection of small structural changes in tissues and organs.

Optoacoustic imaging (OI) is an optical technique that can leverage the advantages of rich optical contrast with a high ultrasonic spatial resolution to capture morphological, functional, and molecular information.

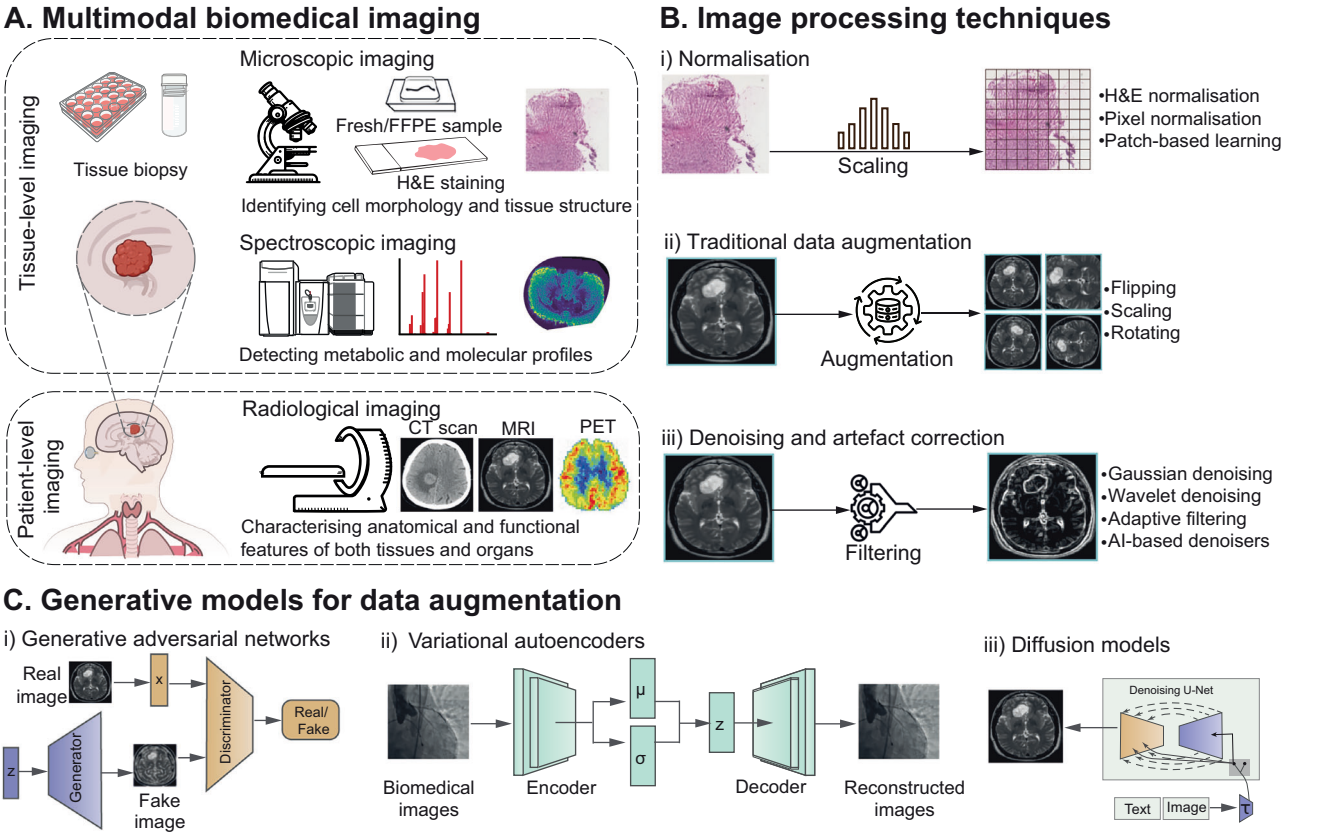

Fig. 1 | Overview of multimodal biomedical imaging and processing techniques. **A** Multimodal biomedical imaging techniques providing complementary biological insights into tissue function and structure, where brain tumour is considered as an example. Microscopic imaging captures the tissue architecture and cell morphology, while spectroscopic imaging detects metabolic and molecular profiles of the tissue. Similarly, radiological images (e.g., Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET)) capture the patient-level image, characterising anatomical and functional features of both tissue and organ. **B** Image processing techniques, including (i) normalisation, which allows for standardising the intensity values across images using Hematoxylin and Eosin (H&E) stain normalisation. Gigapixel images are patched into smaller sizes and patch pixels are normalised; (ii) traditional data augmentation techniques (e.g., flipping, scaling, and rotation), and (iii) denoising and artefact correction, where noise and artefacts from the initial image are removed by applying filtering techniques (e.g., Gaussian- or Wavelet-denoising) or AI-based techniques. **C** Strategies for image augmentation using generative models: (i) generative adversarial networks; (ii) variational autoencoders; and (iii) diffusion models. The human icon was adapted from the NIAID NIH BioArt Source (bioart.niaid.nih.gov/bioart/519). Radiological and histological images were adapted from The Cancer Genome Atlas (TCGA)[176].

Combining microscopic and spectroscopic imaging modalities offers a multidimensional view of tissue health and pathology. For instance, the integration of histopathological and Raman spectroscopy imaging has been shown to enhance cancer diagnosis by associating morphological changes from histopathological images with molecular profiling information from the spectral intensities of Raman images on the same tissue sample[26]. Hence, the integration of histopathological and spectroscopic imaging can overcome the limitations of each modality, providing complementary information for precise molecular-level analysis. Recently, due to its spatially resolved multi-omics molecular mapping (e.g., metabolomics and proteomics), MSI has been increasingly integrated with other biomedical imaging (e.g., immunofluorescent and histopathological) and has shown great potential in identifying biomarkers and investigating tumour heterogeneity at cellular and subcellular levels[6]. Integrating the metabolomics information from MSI and immunophenotype information from imaging mass cytometry has been shown to provide better insights into metabolic heterogeneity across several cell types (e.g., immune cells and cancer cells)[27]. Furthermore, the integration of microscopic and spectroscopic imaging with radiological imaging also has the potential to provide a comprehensive view of molecular information with abnormal tumour regions for precise diagnosis at the patient level[28].

Finally, optical imaging techniques have also been used increasingly to investigate the morphological, functional, and molecular information at the tissue and subcellular levels. The integration of OCT and Optoacoustic Imaging (OI) with spectroscopic imaging (e.g., Raman and multi-spectral) complements the structural imaging with biochemical information of the tissue[29]. Additionally, the combined analysis of OI with radiological imaging and other biomedical imaging has been widely applied in guidance for biopsy, improving tumour phenotype characterisation[30,31]. Specifically, the fusion of OI and MRI has allowed the enhancement of the visibility of blood vessels, including oxygenation levels and haemoglobin concentration from OI, to provide tumour boundary and vascular network details from MRI simultaneously. Such integration offers both structural and functional insights and provides a more comprehensive visualisation of the tumour microenvironment[31].

By leveraging the complementary information from each biomedical imaging (e.g., structural, functional, and molecular aspects), the integration of multimodal biomedical imaging within an AI framework can provide a comprehensive understanding of disease phenotypes and biological processes. However, before integrating multimodal biomedical images, specific preprocessing techniques must be applied to enhance image quality, normalise data, and reduce noise, which successively improve the accuracy and reliability of AI models, as discussed in the following section.

AI-driven biomedical image processing methods

In multimodal biomedical imaging integration, image processing is the fundamental component required to ensure data quality, reliability, and compatibility across diverse imaging sources. Each biomedical imaging

Table 1 | AI-driven processing techniques for biomedical imaging

Processing Method	Model Name	AI Method	Imaging Modality	Target Application	Ref
Artefact correction	HM-EDM	DPM	CT	Motion artefact correction in portable head CT scans.	46
Denoising	Noise2Void	CNN	PET	Unsupervised denoising of PET brain images to improve image quality and quantitative accuracy, especially under low-dose or short-scan conditions.	39
Denoising	CoCoDiff	DPM	CT	Denoising of low-dose CT images using a residual-based diffusion model and contextual slice information.	35
Denoising	MFG-Diff	DPM	PET	Low-count PET image denoising using MRI-guided multimodal feature fusion; enabling SPET-quality image generation with reduced radiation dose while preserving diagnostic fidelity and anatomical detail.	36
Denoising	-	GAN	Ultrasound	Real-time denoising of ultrasound images across anatomical districts; enhancing edge preservation and reducing speckle noise to improve visual assessment	33
Denoising	PADS-Net	GAN	Transcranial ultrasound	Simultaneous denoising and segmentation of midbrain in transcranial ultrasound.	40
Denoising	DeepTFormer	Transformer-based	Mammography	Denoising of mammogram images to enhance visualisation of subtle features such as microcalcifications.	38
Data augmentation	RED-CNN	CNN	CT	Denoising of low-dose CT images using residual learning and deep convolutional-deconvolutional architecture.	34
Data augmentation	-	GAN	MRI	Motion artefact correction and severity assessment in brain MRI; improving image clarity and diagnostic accuracy without requiring paired training data.	43
Data augmentation	MedGAN	GAN	Dermatoscopic/ Macroscopic	Medical image synthesis and augmentation for few-shot learning; supporting lesion localisation and disease classification in scenarios with limited annotated data.	44
Data augmentation	TumorGANet	GAN	MRI	Brain tumour classification using augmented MRI data; improving accuracy and generalisation by addressing dataset imbalance.	49
Data augmentation	TMP-GAN	GAN	Mammography, CT	Pancreatic lesion detection through classification of augmented datasets.	50
Data augmentation	Counter-Synth	GAN	MRI	Gender and age prediction using augmented datasets to address demographic imbalances; improving model fairness and generalisability.	51
Data augmentation	PGGAN, MUNIT	GAN	MRI	Tumour detection in high-grade glioma patients, enhancing classification performance through synthetic data generation.	54
Data augmentation	-	GAN	Histopathology	High-fidelity histopathological image synthesis; expanding rare disease datasets and improving diagnostic classifier performance via realistic tissue augmentation.	55
Data augmentation	DR-VAE	VAE	MRI	Modelling brain states through classification of augmented data.	56

The table reports the most recent papers applying AI-based techniques for biomedical imaging processing. For each paper, the table includes the processing method used, the name of the proposed model (if available), the main AI architecture applied, the image modalities used, a brief description of the target application (e.g., tumour detection, lesion segmentation, or predictive modelling), and the reference. AI Methods: *CNN* convolutional neural networks, *DPM* diffusion probabilistic models, *GAN* generative adversarial network, *VAE* variational autoencoder. Imaging Modalities: *CT* computed tomography, *MRI* magnetic resonance imaging, *PET* positron emission tomography.

modality is characterised by distinct spatial resolutions, contrast mechanisms, and artefactual noise patterns. Hence, integrating heterogeneous imaging data requires robust processing techniques that can normalise imaging features and suppress noise and artefacts.

Radiological imaging modalities such as MRI, CT, and PET inherently contain noise artefacts due to image acquisition parameters, devices, and patient-motion artefacts. These also include non-standardised intensity scales, modality-specific artefacts (e.g., motion blur in MRI, metal streak artefacts in CT), and spatial resolution mismatches across scanners or imaging protocols. Image processing is therefore required for ensuring consistency, reliability, and quantitative interpretability across imaging modalities. Furthermore, microscopic images like tissue Whole Slide Images (WSIs) are high-resolution gigapixel images and cannot be directly integrated with other imaging modalities due to variation in image resolution, memory and computational constraints. To address this, WSIs are typically split into smaller, more manageable image patches. Processing steps, such as tiling, background filtering, and the removal of artefact-laden or out-of-focus regions help ensure that only high-quality, informative tissue areas are retained for multimodal integration and downstream analysis[32].

Traditional image processing techniques, such as normalisation, flipping, scaling and rotating (Fig. 1B(i) and (ii)), do not sufficiently resolve critical degradations in image quality, such as noise, artefacts, tissue folding, or device-specific distortions. These quality degradations can impair multimodal integration and analysis, potentially leading to clinical misinterpretations. Therefore, a deeper exploration of AI-driven processing techniques is required when integrating multimodal biomedical imaging. In this section, we explore how several AI-driven techniques are applied in processing biomedical images (Fig. 1B, C) to improve image quality and support data augmentation. Table 1 reports the latest papers applying AI-driven processing techniques in biomedical imaging, emphasising the growing role of AI in improving biomedical image quality and usability.

Biomedical image denoising and artefact correction

Biomedical imaging quality is crucial to accurately capture biologically meaningful information across different tissues, organs, and pathological states. However, the quality of biomedical images is often compromised by motion artefacts arising, for example, from intricate respiratory patterns and involuntary movements, or device noise. Image noise and artefacts impact multimodal model performance, downstream analyses, and clinical

interpretation, particularly in tasks involving image fusion or quantitative biomarker extraction. Enhancing spatial resolution and reducing noise allows for improved delineation of anatomical structures, such as cortical folds in neuroimaging, microvasculature in retinal imaging, or tumour boundaries in oncology.

Traditional image denoising methods, including median filtering, Gaussian filtering, and total variation denoising, rely on mathematical assumptions and manually crafted priors, Fig. 1B(iii). These methods are sensitive to hyperparameters and often face challenges with modality-specific noise patterns and artefact types inherent to different imaging modalities. In contrast, deep learning models offer content and modality-aware image enhancement by learning complex mappings from noisy inputs to clean outputs directly from data while preserving biologically meaningful features[33]. Supervised deep learning, such as CNN-based models like RED-CNN[34], and perceptual frameworks using GANs, have shown strong performance in biomedical denoising and artefact correction[35,36]. Furthermore, transformer-based models like DeepTFormer, TED-net, and TransCT improve traditional denoising approaches by leveraging global context through attention mechanisms[37,38].

Despite their widespread use, these supervised methods require a large amount of labelled data. To address this limitation, self-supervised and unsupervised methods such as Noise2Void and CycleGAN have emerged, allowing models to learn denoising directly from noisy or unpaired data by leveraging blind-spot networks and domain translation[39,40]. Additionally, generative models like VAEs and diffusion models have advanced denoising and artefact correction capabilities. VAEs, for example, learn structured latent representations that capture essential image features, enabling the reconstruction of clean images by filtering out noise. Denoising Diffusion Probabilistic Models (DDPMs), on the other hand, simulate a gradual noise-removal process through iterative steps, reversing a noise-adding sequence to produce high-quality images with intricate details preserved. Moreover, diffusion models, such as DiffusionMBIR and Deep Diffusion Image Prior, have also demonstrated strong capability for solving inverse problems in biomedical image restoration and reconstruction, where they function as unsupervised priors[41,42].

Besides diffusion models, in artefact correction, such as mitigating motion artefacts in MRI or enhancing low-dose CT scans, GANs-based methods like CycleGAN and MedGAN excel by training on paired datasets to map corrupted images to their artefact-free counterparts, with cycle-consistency ensuring fidelity in critical details[43,44]. Similarly, diffusion-based models, such as the conditional diffusion model for portable head CT, iteratively refine images by modelling noise distributions, demonstrating improvements in motion artefact correction[45,46].

Generative strategies for data augmentation

The generation of synthetic biomedical data, particularly data augmentation, has emerged as a crucial strategy to address limited data availability in multimodal imaging pipelines. In real-world clinical contexts, complete multimodal datasets are often unavailable due to cost, patient-related constraints, or imaging heterogeneity. Traditional augmentation techniques, such as flipping, rotation, scaling, and cropping, have been widely used to introduce statistical variability and increase diversity[47]. However, while effective in many applications, these methods fall short in specificity, especially in multimodal applications, where every modality must be handled with attention to detail, and where tissue-specific patterns or relationships across imaging types need to be taken into account.

To address these challenges, recent AI-based generative models such as GANs and VAEs have shown strong potential to augment the data using available modalities, while incorporating clinical information[48]. GANs (Fig. 1C(i)) have shown strong potential in both augmenting limited datasets and inferring missing modalities in multimodal biomedical imaging. By generating anatomically plausible synthetic images, GAN-augmented data and imputed modalities can enhance training robustness, improve classification performance, and expand sample diversity, particularly for under-represented conditions such as rare tumours[49–51]. Recent methods, including Unified multimodal Image Synthesis, have demonstrated the ability to generate any missing MRI modality from available ones using a shared-discrepant encoding framework[52]. However, challenges such as mode collapse, non-convergence, and training instability can compromise image diversity and clinical reliability, especially in applications requiring fine-grained anatomical accuracy[53,54]. To mitigate these issues, recent works have explored architectural improvements, regularisation strategies, and hybrid model designs[53,55]. Despite these advances, further validation is needed to ensure that generated images support consistent and interpretable outcomes in clinical workflows.

VAEs, on the other hand, are commonly employed in biomedical image processing for synthetic data augmentation (Fig. 1C(ii)). By learning latent distributions from existing imaging data, VAEs can generate anatomically plausible samples to improve model training for tasks such as segmentation and classification[56,57]. However, compared to GANs and transformers, VAEs pose challenges such as blurry outputs, weak cross-modal fidelity, and latent space entanglement[58]. Enhancements like auto-regressive decoders, VAE-GAN hybrids, and supervised disentanglement strategies have been proposed to address these issues, but further work is needed to make them viable for clinical-grade inference across heterogeneous modalities[59].

Beyond GANs and VAEs, other deep-learning frameworks have also been developed to advance data augmentation in biomedical imaging. These include diffusion models (Fig. 1C(iii)), multimodal autoencoders, transformers, and hybrid encoder-decoder architectures that learn shared feature representations from partial modality inputs. Particularly, latent Diffusion Models (LDMs) can effectively capture long-range dependencies, enabling the generation of realistic MRI synthesis[60]. These synthetic datasets are widely applied for addressing class imbalance and simulating rare disease cases. Despite their promising performance, translating these AI-driven approaches to the biomedical domain remains a challenge due to a lack of standardised imaging process, domain shifts, and protocol variability.

Data-driven imaging techniques

Beyond synthetic data generation, AI is widely used in data-driven imaging techniques, representing a significant advancement in biomedical research by enabling deeper insights into disease mechanism. Data-driven imaging techniques extract deep learning and imaging-derived features from high-dimensional biomedical images to provide quantitative information and reveal underlying biological patterns that are not easily visually captured. Early convolutional layers may act as edge or texture detectors, while deeper layers capture more abstract representations such as tumour morphology or tissue patterns. Pretrained models (e.g., ResNet, EfficientNet) and foundation models (e.g., BEPH[61] and RadFM[62]) have been increasingly applied on biomedical imaging datasets to extract feature representations for several downstream tasks. These learning-based feature representations have proven more robust to variations in acquisition settings, noise, and patient anatomy.

Radiomics and radiogenomics derived from data-driven imaging techniques provide detailed insights into disease phenotype and associated genetic characteristics. For example, radiomics can derive tumour information (e.g., tumour texture, shape, and intensity) from radiological images, and radiogenomics correlates radiomics features with genetic profiles of the imaged tissue. These data-driven biomedical imaging offer a quantitative view of tumour phenotype and molecular biology, which can be integrated with the spatial and morphological information from other biomedical imaging to provide a holistic understanding of the disease. PyRadiomics[63] is a widely adopted framework that provides standardised feature definitions and extraction pipelines, enabling reproducibility and facilitating large-scale radiomics studies. With the support of AI, radiomics and radiogenomics are now widely integrated with several biomedical imaging modalities (e.g., radiological and histopathological images) to visualise and characterise tumour tissues and improve precise diagnosis, prognosis, and therapeutic responses[64–66].

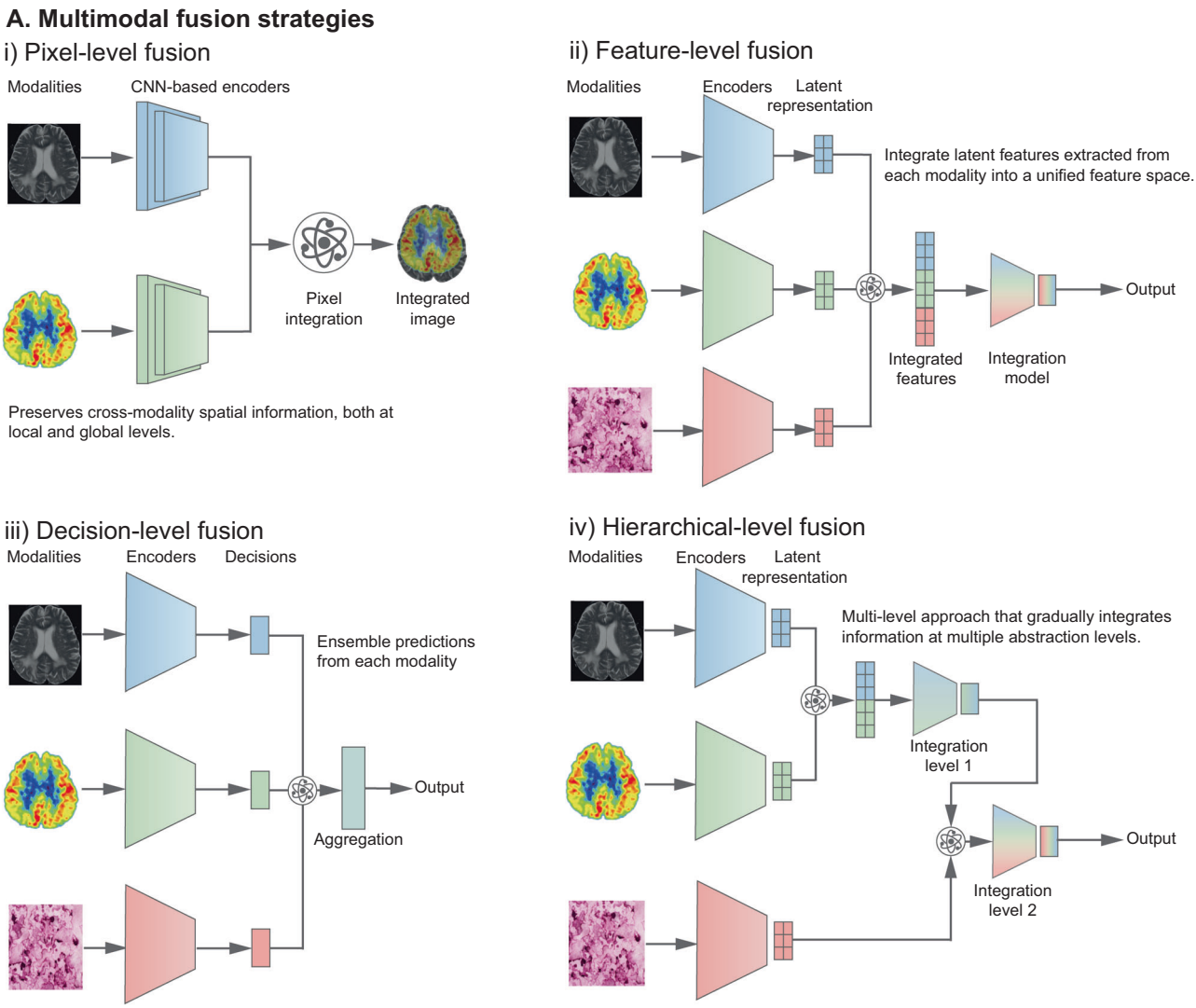

Fig. 2 | Multimodal AI fusion strategies and applications in biomedical imaging. **A** AI fusion strategies to combine and analyse multimodal biomedical images. (i) Pixel-level fusion: raw pixel intensity values from different imaging modalities (e.g., MRI and PET) are combined to generate superimposed images; (ii) Feature-level fusion: latent representations from each modality are concatenated to form integrated features. These integrated latent representations are then fed into a neural network for downstream analysis; (iii) Decision-level fusion: the predictions from each modality are ensembled using majority voting or weighted averaging approaches to generate the final prediction; (iv) Hierarchical-level fusion: the latent representations from each modality are integrated at multiple abstraction levels. **B** Downstream tasks and biological insights. The multimodal integrated architecture can be applied for estimating (i) patients' survival or risk group stratification, and (ii) classification tasks, including subtype or cancer-stage classification. (iii) Explainability techniques (e.g., SHapley Additive exPlanations (SHAP), Gradient-weighted Class Activation Mapping (Grad-CAM) and attention maps) can then be applied to highlight important regions across multimodal images. (iv) Finally, tissue analysis can be performed for cell type annotation or cell segmentation tasks. Radiological and histological images were adapted from TCGA[176].

Integrating AI with multimodal biomedical imaging

The integration of multimodal biomedical imaging within an AI framework is revolutionising disease diagnosis, prognosis, and treatment planning[67]. This section discusses the main AI-based data fusion strategies (e.g., pixel level, feature level, decision level, and hierarchical level) across different biomedical imaging modalities (e.g., radiological, microscopic, spectroscopic, and optical imaging). In particular, we emphasise how different fusion approaches are used in a multimodal AI framework (Fig. 2A) to facilitate a variety of downstream tasks, including risk prediction, cancer detection, cell type identification and annotation (Fig. 2B). Lastly, Table 2 reports the recent multimodal AI applications in biomedical imaging, showcasing advancements in disease diagnosis, prognosis, and tumour segmentation, highlighting the role of multimodal AI for biomedical imaging in healthcare and medical research.

Table 2 | Recent multimodal AI fusion strategies applications in biomedical imaging, listed by fusion technique, pixel-level, feature-level, decision-level, and hierarchical-level fusions

Integration techniques	Imaging modalities	Downstream tasks	Anatomical sites	Ref
Pixel-level	MRI, PET	Preserving imaging details from both modalities to enhance clinical diagnosis	Brain	70
	MRI, PET	Enhancing image quality and visualisation (higher brightness and clearer edges) to improve the clinical diagnosis	Brain	71
	MRI (T1, T2), PET, SPECT	Preserving imaging details from both modalities to enhance clinical diagnosis	Brain	72
	MRI, PET	Avoiding blurring and reducing visible texture detail loss to improve the clinical diagnosis	Brain	76
	MRI, SPECT	Improve radiological image characterisation by combining anatomical soft-tissue structures with functional metabolic features in brain imaging	Brain	77,78
	MRI, PET	Alzheimer's disease detection	Brain	73
Feature-level	OCTA, OCT (structure), ultrasound (B-scan flow)	Age-related macular degeneration diagnosis	Eyes	84
	Raman, FTIR spectroscopy	Thyroid tumour diagnosis	Cervical region	82
	MRI and fMRI	Schizophrenia diagnosis	Brain	81
	MRI (T1, T1c, T2, FLAIR)	Brain tumour	Brain	80
	Raman, FTIR spectroscopy	Cancer diagnosis (e.g., lung cancer, glioma, and thyroid)	Lung, brain, neck	107
	WSI, MRI	Mutation status prediction	Brain	28
	Mammography, ultrasound	Breast cancer subtype identification	Breast	86
	MSI, WSI	Lung cancer subtyping area segmentation	Lung	111
	Mammography, ultrasound	Breast cancer screening	Breast	83
Decision-level	Radiomics (extracted from CT), WSI	Ovarian cancer risk stratification	Ovary	65
	Microscopic imaging (AFMI, BFMI, and OPMI)	Breast cancer diagnosis	Breast	87
	Radiomics (extracted from CT), WSI	Prostate cancer progression risk prediction	Prostate gland	88
	MRI, WSI	Brain tumour subtype identification	Brain	85
	MRI (T1, T2, DCE)	Breast cancer diagnosis	Breast	89
	MRI (T1, T1c, T2, FLAIR)	Brain tumour subtype identification	Brain	90
	LIBS and Raman	Lung cancer stage prediction	Lung	110
Hierarchical-level	MRI (T1, T1c, T2, FLAIR)	Brain tumour segmentation	Brain	94,102
	CT, WSI	Gastric cancer diagnosis	Stomach	92
	CT, MRI	Brain image quality enhancement	Brain	93
	Fundus retinal photography, OCT, fluorescein angiography	Treatment-requiring retinal vascular disease detection	Eyes	114
	CT, MRI, endoscopy	Med-SAM, a foundation model for universal medical image segmentation	Multiple	8

The table also reports the imaging modalities used in each study, the downstream tasks, the anatomical site, and the corresponding reference. *AFMI* autofluorescence multispectral imaging, *BFMI* backgrounded membrane imaging, *CT* computed tomography, *fMRI* functional magnetic resonance imaging, *FTIR* fourier transform infrared, *LIBS* laser-induced breakdown spectroscopy, *MRI* magnetic resonance imaging (T1, T1-weighted imaging with contrast (T1c), T2, *FLAIR* fluid-attenuated inversion recovery, *DCE* dynamic contrast-enhanced are different types of MRI sequences), *MSI* mass spectrometry imaging, *OCT* optical coherence tomography, *OCTA* optical coherence tomography angiography, *OPMI* operating microscope imaging, *PET* positron emission tomography, *SPECT* single-photon emission computed tomography, *WSI* whole slide imaging.

Integrating different imaging modalities: data fusion techniques

Data fusion strategies in biomedical imaging integrate multiple modalities to enhance the understanding of anatomical structures and physiological processes, thereby improving disease diagnosis, prognosis and therapeutic monitoring[67]. These strategies, including pixel-level, feature-level, decision-level, and hierarchical-level fusion approaches (schematically represented in Fig. 2A), can leverage the complementary information that captures structural, functional, and molecular aspects of diseases to improve the predictive power of the AI model and mitigate individual modality limitations.

Pixel-level fusion. Pixel-level fusion (also referred to as early fusion, Fig. 2A(i)) combines different biomedical imaging modalities of the same sample region at a pixel-level to create a single and enhanced image[68]. Two main techniques, including spatial domain and frequency domain fusion (i.e., multi-scale-based transform), are often used in pixel-level fusion to improve image quality and generate less spatial distortion[69]. Spatial domain fusion utilises basic pixel-level techniques, such as average/weighted average, maximum, principal component analysis, and gradient filtering, to fuse biomedical images. On the other hand, frequency domain fusion (e.g., multi-scale transformation and pyramidal methods) aims to fuse the frequency quantities extracted from several biomedical images via Fourier transform or deep learning approaches, followed by inverse transformation to obtain the final fused image. For instance, combining pixel-level information from CT, MRI, and PET has been shown to enhance the spatial and anatomical features of biomedical

images.[70–73]. Such integration not only preserves the original information from each original modality (e.g., tissue structure) but also enhances complementary insight and interrelations across modalities[74]. As a result, pixel-level fusion can improve the image quality (e.g., colour brightness, edge and local features) compared to using a single modality only[71,75].

Deep learning-based approaches (e.g., CNNs, GANs, and transformers) have also been increasingly used to transform the spatial or signal information of each modality through pixel-level fusion[76–78]. For instance, in pixel-level fusion, CNNs have been employed to capture spatial hierarchies from MRI and SPECT images, while transformer-based approaches and attention mechanisms have been applied to integrate these features and generate a more accurate representation of anatomical and functional tissue aspects[77].

Pixel-level fusion has proven effective in enhancing the visualisation quality, segmentation, and characterisation of biomedical imaging, providing more informative images and facilitating better disease diagnosis and screening[68]. By integrating complementary information from different imaging modalities of the same body region (e.g., the brain), pixel-level fusion can combine functional information (e.g., metabolic activity from PET and SPECT) with precise anatomical detail (e.g., tissue structure from MRI and CT) to generate a clearer and more comprehensive image, thereby supporting better-informed treatment decisions[23,70]. However, pixel-level fusion combines raw images, and it can only be applied to imaging representing the same sample region. Hence, this fusion technique is sensitive to image quality and misalignment between images. For example, a misalignment between modalities such as MRI and PET will result in artefacts (e.g., blurring and noise amplification) and structural distortions, reducing the diagnostic quality of the output. Therefore, when noise and misalignment are present in imaging modalities, additional preprocessing, such as image registration and denoising, is required before integration (as discussed in Section "Biomedical image denoising and artefact correction"). Registration error can also be applied to determine the spatially matched region locally (e.g., regions of interest) or globally (e.g., entire image section) before pixel-level fusion[79].

Feature-level fusion. Feature-level fusion (also referred to as intermediate or joint fusion, Fig. 2A(ii)), in contrast, can overcome the challenges of pixel-level fusion by using deep learning models to extract meaningful features or latent representations from each modality before integrating them for downstream analysis[67]. By leveraging extracted features rather than raw pixel data from each modality, feature-level fusion has demonstrated greater robustness in handling misalignment and noise compared to pixel-level fusion[67]. The modality-specific noise, redundant information, and imaging quality variability are reduced during the extraction of the latent representation using advanced deep-learning-based approaches. In particular, CNNs, RNNs, VAE, and vision transformers are often applied to extract latent representation in feature-level fusion to enhance interpretability and predictive performance of multimodal AI[80–83].

Additionally, feature-level fusion also offers more flexibility in the regions of interest to be integrated. Unlike pixel-level fusion, which requires exact pixel alignment between input images, feature-level fusion allows for the integration of features extracted from different regions, enabling a more holistic understanding of the patient's condition. For instance, feature-level fusion of different biomedical imaging, such as optical imaging (e.g., OCT and OCT angiography), radiological imaging (e.g., ultrasound and MRI) and spectroscopic imaging (e.g., Raman and infrared spectroscopy), has demonstrated improvement in disease diagnosis[82,84]. Histopathological imaging has been integrated with multi-sequence MRI scans through feature-level fusion, combining information from multiple views, including molecular-level and cellular-level (from histopathological imaging), and texture-level information (from MRI). Such integration has provided a more comprehensive perspective and enhanced clinical diagnosis[28].

Feature-level fusion has been successfully applied in many biomedical tasks since it is more robust to noise and image misregistration compared to pixel-level integration. However, due to the complexity of the AI approach used for fusing the features from different modalities, feature-level fusion often requires a complex architecture, and it is computationally expensive, especially when working with high-resolution biomedical images. Additionally, the choice between integrating the same or different regions depends on the nature of the dataset and clinical objective. While same region feature-level fusion enhances the diagnostic accuracy and interpretability of the multimodal AI, different region fusion supports a more comprehensive, systemic understanding of disease processes.

Decision-level fusion. To overcome the limitations of feature-level fusion, in decision-level fusion (also referred to as late fusion, Fig. 2A(iii)), the decisions or predictions from each modality are combined using mathematical operations, such as majority voting or weighted average, to generate the final outputs[67]. Moreover, by combining high-level decisions inferred from individual modalities, this fusion strategy has shown robust results in many applications since it can handle data misalignment and noise between multiple imaging modalities. Decision-level fusion can integrate predictions from different biomedical imaging modalities, such as radiological (e.g., MRI, ultrasound, and mammography), radiomics, and microscopic imaging (e.g., WSI) to generate a more robust and accurate disease diagnosis and improve subtype identification[65,85–88]. It has also been applied to fuse different MRI perfusion techniques (e.g., T1, T2, and dynamic contrast-enhanced) and improve breast cancer detection[89,90]. Moreover, decision-level fusion has also improved cancer patients' outcomes by integrating biomedical imaging from different biological views, such as radiomics from CT scans and histopathological features from WSI[65]. CNNs, especially pre-trained networks (e.g., ResNet), have also been widely used in decision-level fusion to generate predictions from each biomedical imaging modality[89,90]. Due to the flexibility in the selection of modality-specific architecture in decision-level fusion, foundation models are also increasingly applied to enhance the accuracy, robustness, and generalisability of multimodal approaches[8,91].

Decision-level fusion is promising for handling high-resolution images and generating more robust predictions. Similar to feature-level fusion, it can also deal with heterogeneous data and preserve the modality-specific features, facilitating the integration of different biological organisations. Hence, decision-level fusion is commonly used when biomedical imaging originates from different regions, particularly when integrating multi-level imaging (e.g., patient-level, tissue-level, and molecular-level). However, since the input data is not integrated directly, but the decision-level fusion combines a higher level of abstraction (i.e., predictions), the detailed information (e.g., morphological and functional) present in the original images might be lost. Moreover, the running time of decision-level fusion approaches might be impacted by the image resolution of each modality. For instance, due to the high-resolution nature of most biomedical imaging, patching-based techniques are required when working with microscopic imaging (e.g., WSI), while this is not required for radiological imaging, introducing further challenges when applying decision-level fusion techniques.

Hierarchical-level fusion. To mitigate the limitation of multiple-resolution image integration, hierarchical-level fusion (Fig. 2A(iv)) combines multiple abstraction levels (e.g., pixel-level, feature-level, and decision-level) to improve image quality and increase the accuracy and robustness of the resulting AI model[67]. By integrating information at different levels from pixel- to decision-level, hierarchical fusion can combine the strengths of different fusion types, providing a more comprehensive understanding of the disease and biological processes. Hierarchical-level fusion is often applied to integrate gigapixel imaging (e.g., histopathological imaging) and radiological images (e.g., MRI and CT) to improve survival prediction and interpretability of AI models[92]. It can also handle biomedical imaging with huge differences in spatial scales between patch-level and region-level on WSI and tissue-level on MRI[92].

Hierarchical-level fusion is particularly effective in capturing spatial dependencies in modalities, such as MRI and CT, which involve three-dimensional spatial information and require specialised processing pipelines and architectures tailored to their unique features. Additionally, a multi-scale hierarchical fusion approach that combines lower-scale features (e.g., pixel intensities, local details, and texture information) and higher-scale features (e.g., global structures) has been recently applied to integrate CT and MRI and generate better-fused images in terms of image quality and richer details and textures[93]. CNNs are commonly used in hierarchical-level fusion to extract and compute features, while vision transformers and attention mechanisms have been recently applied to capture long-range dependencies within high-resolution images, leading to more accurate and interpretable results[92–94].

Hierarchical-level fusion remains a promising approach since it can overcome the limitations of other fusion approaches (i.e., it is robust to noise and it can deal with large differences in spatial scales and image resolutions). It also provides more flexibility in designing and optimising fusion strategies by enabling the selection and combination of various fusion strategies at different levels. Despite its promising performance, combining multiple abstraction levels requires complex architectures, resulting in higher computational costs compared to other fusion strategies. Although tailored approaches are applied to accommodate the differences in scale and resolution (e.g., Mixture of Experts (MoEs), where modality-specific expert networks are assigned to process different imaging modalities[95]), achieving effective optimisation and information fusion across heterogeneous modalities in hierarchical-level fusion remains a challenge. Additionally, when misalignment and noise are present in the imaging modalities, additional preprocessing (e.g., image registration and resizing) and multiple transformation steps are required before and during the integration. Therefore, although hierarchical-level fusion offers improvements in model performance and flexibility, it still requires careful consideration of the trade-offs between model performance and complexity[67].

Downstream tasks and biological insights

The integration of multimodal imaging within an AI framework has been increasingly used to enhance diagnosis, prognosis, subtype identification, treatment response, and understanding of tissue heterogeneity, as schematically represented in Fig. 2B.

Different types of radiological images (e.g., mammography, MRI, and CT) have been recently integrated with multimodal AI to improve sensitivity in cancer risk assessment, and enhance the accuracy of tumour and lesion detection used for preoperative planning and prediction of metastasis[96–99]. Integrating complementary information from multiple sequencing techniques in cardiac magnetic resonance (e.g., auxiliary and late gadolinium enhancement) has also improved the segmentation results and reliability compared to using one imaging modality only[100,101].

Recently, transformer-based models, such as Swin UNETR, incorporate self-attention mechanisms to efficiently capture long-range dependencies and multi-scale contextual features, enhancing tumour segmentation accuracy from multimodal radiological images (e.g., MRI) and multi-scale structures often observed in tumour regions[102,103]. These models are particularly effective for MRI due to their ability to model complex anatomical structures. Also, PET and SPECT scans integrations have been successfully applied to enhance metabolic and anatomical analysis and provide a richer perspective in complex cases such as the analysis of head and neck cancer[23]. Other commonly applied combination approaches are based on CT and MRI images, which contribute to high-resolution anatomical details. In fact, CT and MRI are often combined with PET to provide complementary insights into metabolic activities (e.g., glucose uptake) from PET and anatomical details from CT and MRI and improve diagnostic and prognostic accuracy[22]. Saliency maps have been used to guide the fusion process, enhancing the explainability of multimodal biomedical imaging by ensuring that the fused images highlight visually and clinically relevant features[70]. The synergy across radiological imaging modalities has been shown to improve predictive capabilities and has the potential to further enhance AI systems' relevance in healthcare[104,105].

The integration of different types of microscopic and spectroscopic imaging has also provided the opportunity to gain more comprehensive information by correlating structural (from microscopic imaging) and functional information (spectroscopic imaging), thereby leading to more accurate diagnosis, subtype, and biomarker identification[26,106,107]. For example, multimodal AI integrating digital histopathological and Raman spectroscopy imaging has shown improvement in cancer diagnostic accuracy compared with unimodal AI[26]. Besides, histopathological imaging has also been integrated with radiological imaging (e.g., MRI) to improve the robustness and predictive performance of multimodal AI models in predicting mutation status and classifying cancer subtypes[28,108]. Additionally, Raman spectroscopy imaging is a common modality integrated with other biomedical imaging, such as infrared spectroscopy and fluorescence imaging, to enhance disease diagnosis, biomarker detection, and better quantify the metabolic activities of different cancer subtypes[82,106,109,110]. Moreover, hyperspectral imaging and MALDI MSI offer the potential to more precisely delineate tumour margins by exploiting spectral contrasts between healthy and diseased tissue and enabling the generation of virtual histological stains to reduce the cost and variability of conventional chemical staining and conserve precious tissue samples[111,112].

The advancement of AI and data integration methods has also facilitated the integration of high-resolution imaging, such as optical and data-driven biomedical imaging. Their integration has shown great potential in better characterising disease, enhancing the risk stratification of patients, accurate diagnosis, and identifying biomarkers and investigating disease[6,31,65,113]. For example, by integrating OI and OCT with other radiological, microscopic, and spectroscopic imaging modalities, including MRI, X-ray, Raman spectroscopy, and fluorescence imaging, it has been possible to leverage the complementary information from each modality and enhance clinical cancer diagnosis and therapeutic monitoring[31,113]. Moreover, OCT and OI have also been integrated with other ophthalmic imaging methods, such as fundus photography and fluorescein angiography, to facilitate AI-driven diagnosis of complex retinal diseases[114]. OCT-derived structural features, including the retinal nerve fibre layer and the ganglion cell-inner plexiform layer, have also been used to predict neurodegenerative conditions such as Alzheimer's disease and mild cognitive impairment, highlighting the role of multimodal AI to enhance the precision of disease diagnosis[115].

Although multimodal biomedical imaging provides a promising solution to enhance clinical decisions, translation of these AI models often faces several challenges related to scalability, cross-modality generalisation, and adaptability to complex clinical scenarios. Foundation models like MedSAM[8], BiomedParse[116], and BiomedGPT[12], further push the boundaries of multimodal applications by offering generalisability and robust performance across multiple imaging modalities, outperforming traditional and specialist models in accuracy and adaptability. They leverage pre-trained embeddings and attention mechanisms to adaptively refine segmentation outputs, marking a key step towards universal biomedical segmentation solutions[8].

Finally, to fully exploit the potential of AI in bioimaging, it remains essential to integrate imaging data with other biomedical sources, including omics (e.g., transcriptomics, proteomics, and genomics), EHRs, and time-series data from wearable devices. In fact, cross-modality algorithms enable AI systems to combine these diverse data sources to offer a comprehensive understanding of patient health and disease progression[117,118]. For instance, integrating imaging with genomic data can reveal links between structural abnormalities and molecular profiles, which enhances understanding of disease mechanisms[117,119,120]. This synergy enhances the ability to identify biomarkers that correlate imaging phenotypes with genetic variations. Additionally, linking cardiac MRI with genomic data has identified loci associated with aortic dimensions, offering mechanistic insights into cardiac function that single-modality analyses might miss[121,122]. Similar strategies have been used to improve prognostic models in cancer by connecting gene expression profiles with visual patterns in biomedical images to enable more accurate prediction of recurrence risk in breast cancer patients[123]. Beyond

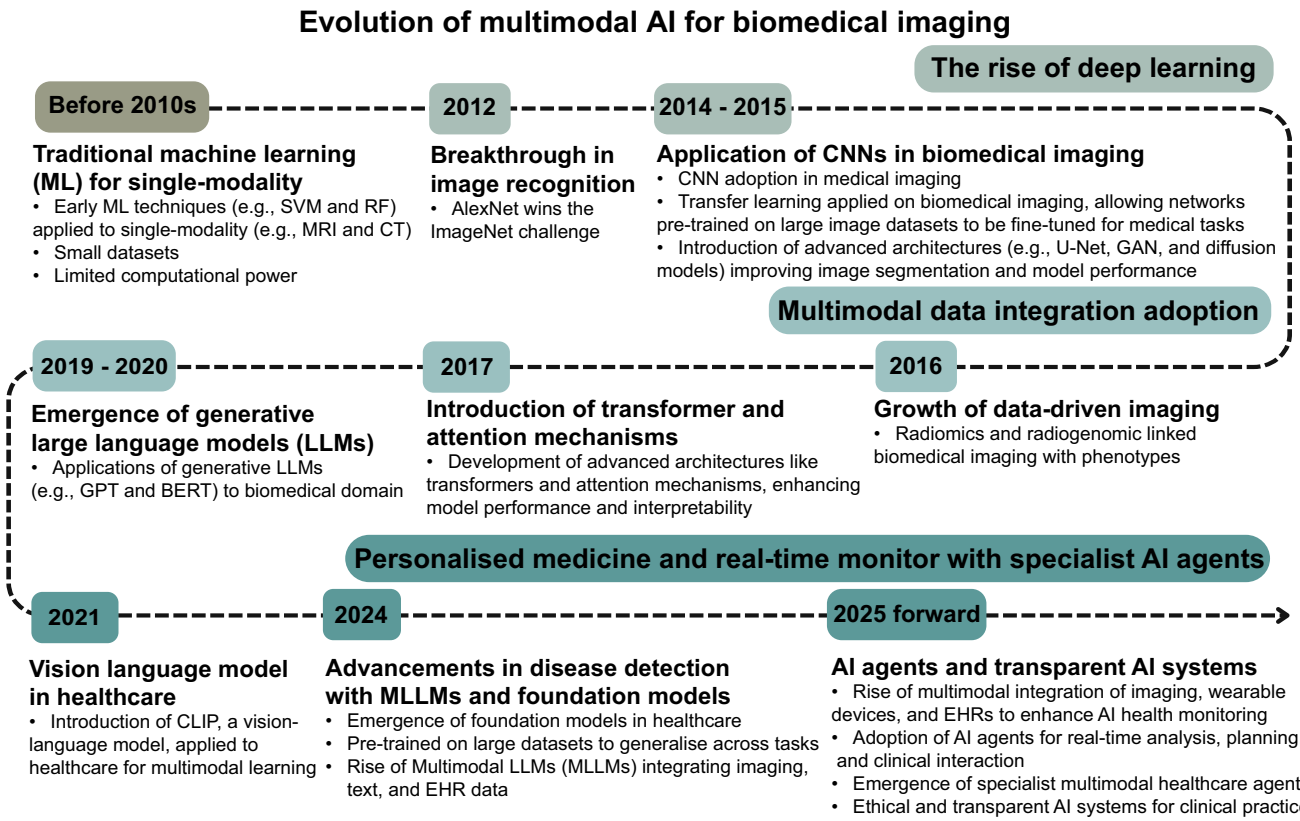

Fig. 3 | The evolution of AI in multimodal biomedical imaging. Early approaches relied on traditional machine learning (e.g., Support Vector Machine (SVM) and Random Forest (RF)) applied to a single imaging modality. The breakthrough in image recognition with the success of AlexNet in the ImageNet challenge has opened a new era of deep learning from 2012, with more applications of CNNs and pre-trained networks on large image datasets to be fine-tuned for biomedical tasks. Since 2016, the field has expanded rapidly toward multimodal integration, driven by the growth of data-driven imaging and the integration of multiple imaging types to gain more biological insights. The emergence of transformers and LLMs have extended AI applications to the biomedical domain with natural language generation and mining. More recently, foundation models and multimodal large language models (MLLMs) have demonstrated the capacity to generalise across several downstream tasks without the need for task-specific architectures. Looking ahead, the rise of AI agents, especially specialist multimodal healthcare agents, is anticipated to streamline workflows.

prognosis, early detection is another key benefit. Since molecular changes often occur before structural abnormalities appear, combining omics with imaging allows for earlier identification of disease or detection of patients at higher risk, even before symptoms are visible[124].

In parallel, EHR data contextualises imaging findings within patient histories and as a result, improves diagnostic relevance, while wearable devices provide real-time monitoring that complements static imaging scans. Together, these modalities complement imaging by adding clinical and functional dimensions and enable a more complete and time-resolved understanding of patient health.

Furthermore, the development of MLLMs enables the integration of biomedical images (e.g., radiological and pathological) with natural language descriptions to enhance biological and clinical insights and improve explainability[125]. Additionally, text integration supports zero-shot learning by enabling the development of generalist models[126]. Integration of text with single imaging modalities, such as histopathology[125,127,128] or radiology[129–131], remains an active area of research and continues to yield important advances. In fact, more recently, there has been increasing interest in frameworks that integrate multiple imaging modalities alongside text[132–134] to enable more comprehensive analyses and deeper insights into complex biological systems.

Evolution of AI in Multimodal Biomedical Imaging

The advances outlined above reflect the AI innovations in biomedical imaging over the past fifteen years, as illustrated in Fig. 3. The field was initially dominated by traditional machine learning approaches (e.g., Support Vector Machine (SVM) and Random Forest (RF)) applied to single-modality imaging data (e.g., MRI, CT and X-rays) in the early 2010s. However, insufficient computational power and the scarcity of large publicly available datasets have limited the performance and applications of AI in biomedical imaging. The success of AlexNet[135] in the ImageNet challenge has opened a new era of applying deep learning, particularly CNNs and pretrained models, to biomedical imaging[136]. Architectures, such as U-Net[137], GANs[138], and diffusion models[139], subsequently revolutionised the applications of AI, enabling models to focus on regions of interest, improve image quality and data augmentation, and thereby enhancing diagnostic precision. The rise of deep learning also fostered the growth of data-driven imaging approaches (e.g., radiomics and radiogenomics) connecting imaging features to disease phenotypes[140] and facilitated advanced analyses in cell and molecular signatures from high-resolution imaging (e.g., microscopic and spectroscopic)[141]. From 2016, the field expanded rapidly toward multimodal integration, incorporating multiple imaging types (e.g., radiological, microscopic imaging and data-driven imaging) to gain more insights into disease mechanisms. The introduction of the transformers and attention mechanisms[142] further advanced the biomedical imaging field, enhancing model performance and interpretability. In parallel, the emergence of generative large language models (LLMs), such as GPT and BERT, extended AI applications to the biomedical domain with natural language generation and mining. Vision-language models in healthcare, like CLIP introduced in 2021[143], have improved generalisability, reduced reliance on manual annotation and opened a new era of personalised medicine and real-time monitoring. More recently, foundation models and MLLMs, trained

on extensive multimodal datasets incorporating biomedical imaging, clinical notes and EHRs, have demonstrated the capacity to generalise across a wide range of downstream tasks (e.g., diagnosis, prognosis and biomarker identification) without the need for task-specific architectures[8,144,145]. Importantly, the field has shifted from a single and isolated imaging modality to a multimodal and patient-centric paradigm, enabling a deeper understanding of underlying biological insights of the disease and supporting the development of personalised medicine.

Looking ahead, the integration of diversified multimodal biomedical imaging offers an opportunity for the rise of new AI systems like AI-agents to integrate AI models with the clinical bioimaging ecosystem, including biomedical databases, diagnostic tools, and analytical software. These agents can streamline workflows by incorporating imaging, wearable devices, and EHRs, and providing real-time analysis, planning and monitoring. These multimodal proactive AI-driven health monitoring systems will not only support real-time analysis and clinical decision-making but also provide tools for planning and continuous interaction between clinicians and AI, which later support the development of specialist multimodal healthcare agents to enhance diagnostic precision and personalised care. Importantly, these AI innovations must be accompanied by a strong emphasis on ethical and transparent AI design, ensuring that clinical practices benefit from systems that are explainable, trustworthy, and aligned with patient safety and regulatory standards.

Challenges in AI and multimodal bioimaging

Although the integration of AI into biomedical imaging has improved diagnostics and treatment planning, several challenges still remain. For example, data variability, noise, and inconsistencies across protocols and demographics must be considered before combining different imaging modalities. The black-box nature of AI also raises concerns about interpretability, clinical trust, and ethical factors, including privacy, security, and bias. This section discusses the key challenges of multimodal AI in bioimaging, from data variability to model interpretability and ethical considerations, while proposing potential solutions and future research directions to address these issues.

Data quality and variability

Biomedical imaging data often come from diverse sources, increasing their variability and affecting the overall dataset quality[146]. Hence, the success of AI applications in biomedical imaging heavily depends on data quality and data variability. Data quality refers to the reliability and suitability of imaging data for AI use. Accurate images and high-quality annotations are essential for building robust AI models[147], since poor-quality data, with noise, artefacts, or inconsistencies, can hinder AI's ability to learn meaningful features and bring about reduced model accuracy and robustness[148]. Similarly, data variability, which represents inherent differences in imaging datasets, arises from patients' diversity, acquisition parameters, and imaging modality structure[149]. As expected, integrating multiple modalities adds complexity, as small sample sizes from paired modalities often fail to represent all conditions, which sequentially leads to biases and limited generalisability[118]. Additionally, temporal changes in patient conditions and treatment responses pose further challenges to maintaining data consistency and ensuring fairness[10,150].

In order to tackle the challenges of data quality and variability, several strategies have been developed. For example, data curation and harmonisation, involving standardised protocols for image acquisition, annotation, and data management, can help minimise variability and improve data quality in medical AI applications[147].

Moreover, missing data modalities pose another challenge when working with multimodal biomedical imaging. Recently, cross-modal data imputation has leveraged generative models and attention mechanisms to synthesise missing modalities from available data[151,152]. In parallel, contrastive learning approaches have been used to align modalities in a shared latent space, allowing models to leverage correlations and similarities across modalities, even when modalities are missing[153]. Leveraging data enhancement techniques (e.g., data augmentation for transforming images and generative AI for synthetic biomedical imaging generation and translation) can also expand the diversity and size of the training datasets, enhancing model robustness when working on noisy biomedical images or integrating multiple modalities across complex biological systems[154,155]. Additionally, developing robust AI algorithms that are inherently resistant to noise and data variability is essential. This can be achieved by adopting transfer learning approaches based on large pre-trained models that capture useful low-level features, or by incorporating data augmentation to simulate biologically relevant perturbations (e.g., variations in staining intensity, imaging artefacts, or cell morphology).

By using multi-task learning strategies that combine diverse biomedical datasets and label types (e.g., cell types, disease states, or pathway activity), or employing techniques that systematically account for batch effects and technical variations across different data sources of multi-platform biomedical datasets can also address the issue of low data quality and variability and develop more accurate, robust, and fair AI models that will lead to improvement in patient care and advance medical research[156].

Interpretability of multimodal AI approaches

The application of multimodal AI in biomedical imaging has demonstrated improvement in predictive performance by combining complementary information across various imaging modalities (e.g., MRI, CT, PET scans, and histopathological images). However, despite their advantages, multimodal frameworks introduce multiple interpretability and explainability challenges. Hence, developing effective interpretability methods remains crucial to uncover biologically and medically relevant regions or insights from biomedical images[157]. To address this challenge, multimodal interpretability techniques can be applied to underline the modality-specific features contributing towards prediction and elucidate whether any modality overshadows the others. Such transparency is necessary to ensure that multimodal frameworks uniformly utilise each modality's unique features, preventing the suppression or overshadowing of potentially important data.

While post-hoc model-agnostic interpretation techniques, such as SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and permutation-based feature importance, have gained popularity due to their flexibility, these techniques are mainly limited to the interpretation of single modality approaches[158]. In fact, these techniques often lack the ability to capture complex interactions between modalities or interpret cross-modality relationships effectively.

Model-specific interpretation techniques, like Grad-CAM (Gradient-weighted Class Activation Mapping), Grad-CAM++, and attention-based mechanisms, have shown promise in interpreting complex models[159]. Model-specific approaches provide precise interpretability by highlighting crucial features within high-dimensional data. For instance, attention layers in CNN-based models can highlight relevant regions in medical imaging, such as MRI scans, allowing these models to generate explanations that align closely with the medical decision-making process[118,160]. While these techniques have been successfully applied in single-modality models, multimodal adaptation brings additional challenges in balancing computational costs with the accuracy of modality interpretations. Moreover, model-specific approaches also struggle with cross-modal interpretations, especially in multimodal frameworks, where diverse data types (e.g., MRI and CT scans) must be interpreted in a unified manner. The complexities of recently developed transformer-based and multimodal models in biomedical imaging present further interpretability challenges, as existing methods often fail to elucidate inter-modality interactions, limiting their effectiveness in leveraging multimodal data[103,161].

Ethical considerations in AI applications

The rapid integration of AI technologies in biomedical imaging brings several ethical challenges, ranging from patient privacy and data security to biases in diagnostic algorithms and the transparency of AI-driven decisions. Given the sensitive nature of medical data, ensuring patient confidentiality is paramount. AI systems often require vast amounts of data to perform accurately, making the use and storage of patient images a critical concern[162].

Particularly, the deployment of emerging techniques in multimodal biomedical imaging, such as foundation and MLLMs, raises privacy concerns. Adherence to data protection regulations such as the EU's General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) in the U.S. is essential. Ethical frameworks must prioritise secure, anonymised handling of patient data to safeguard privacy[163], especially following the COVID-19 pandemic[164].

Additionally, computational resources present another challenge in adopting multimodal biomedical imaging, particularly with the deployment of emerging techniques like MLLMs. These models are resource-intensive, often requiring high-specification hardware or cloud-based processing, which can involve sensitive patient data and necessitate robust governance frameworks to ensure responsible use[145]. Running MLLMs locally demands substantial infrastructure, which limits their adoption in resource-constrained settings like hospitals or research labs. Developing lightweight models could help address these barriers to enable accessibility.

Bias in healthcare and AI algorithms is another critical concern in biomedical imaging. Training models on datasets that lack diversity can lead to biased outcomes, limiting the model's effectiveness across diverse populations. This risk is heightened in biomedical imaging, where underrepresented groups may face disproportionate impacts if the training is done on homogeneous datasets. Hence, ethical AI development must incorporate fairness and mitigation by ensuring representative datasets and rigorous bias assessment[165]. Lastly, accountability in AI systems should be delineated, since cases of diagnostic errors or adverse clinical outcomes need to be assigned clear responsibilities, from the AI developers to the healthcare providers and institutions. Establishing a well-defined accountability framework can aid in resolving issues related to AI-assisted healthcare[166].

Several AI-based approaches have been recently proposed to address the above challenges. In the following section, we present the emerging solutions and propose future directions.

Perspective section: future directions

With increasing access to large, high-resolution datasets, advances in AI are transforming image analysis, enhancing data integration, and providing insights for precision medicine[74,167]. In this section, we highlight the emerging AI techniques that will likely have a strong impact on multimodal bioimaging applications. We briefly describe their role and potential impact on clinical efficiency and personalised care. We also present the impact of the integration of imaging data with other modalities, including omics and electronic health records, to provide a more comprehensive understanding of the disease under investigation.

Emerging techniques in multimodal bioimaging

AI techniques for harmonising different bioimaging datasets have recently been applied to overcome common challenges in resolution, scale, and acquisition methods[68,168]. Specifically, AI enhances dataset alignment and fusion to ensure an accurate integration of complementary data. In direct imaging like PET and MRI, deep learning-based methods have been employed to improve the registration of simultaneously acquired data, while in indirect combinations like Raman spectroscopy with MRI, multimodal AIs have been used to resolve spatial mismatches and enhance co-localisation of heterogeneous data types[74]. The emerging applications of generative AI models can also address the lack of data in some modalities by generating synthetic data or making accurate predictions when only limited information is available[169].

Foundation models and MLLMs are recently emerging as powerful techniques in multimodal bioimaging. These models, trained on extensive and diverse datasets, can perform a multitude of tasks with minimal fine-tuning, making them particularly valuable in bioimaging where labelled data are scarce and costly. By leveraging transfer learning and pretraining on large-scale unlabelled data, foundation models can learn generalisable patterns, which makes them adaptable to various imaging tasks while mitigating the variability introduced by different imaging protocols and sources[170].

MLLMs extend this potential by integrating text-based metadata with imaging data, enabling richer contextual understanding[95]. These models boost the capabilities of foundation models and remove the need for task-specific fine-tuning[171], although their application in biomedical imaging still faces several challenges. Aligning visual features with biomedical language is a primary hurdle, as medical images often lack clear object boundaries or categorical labels, and associated text may be incomplete or inconsistent[172]. This misalignment complicates training, particularly with limited or noisy data, which can hinder the connection between visual tokens and semantic meaning. This semantic gap between imaging data and biomedical terminology requires specialised approaches that can handle the ambiguity in medical image features.

While attention mechanisms and image-text grounding can improve explainability in MLLMs[170], they often lack the precision needed for clinical decision support. To address this limitation, pixel-level aware MLLMs have been recently developed that extend beyond image-level processing to provide more granular feature attribution[133,134]. However, these approaches are still nascent, and their reliability in clinical settings requires further validation. Importantly, the integration of multiple imaging modalities further amplifies the challenges of interpretability, as models must disentangle and attribute features across heterogeneous data sources in a clinically meaningful way.

Looking ahead, MLLMs have the potential to substantially advance bioimaging by enabling integrated analysis and deeper contextual understanding across diverse imaging and textual modalities. Building on this foundation, emerging agentic systems, which use MLLMs as their core reasoning engines, further extend these capabilities through autonomous decision-making, intelligent tool selection, and workflow automation. Nevertheless, key challenges remain in aligning features across modalities, ensuring clinical interpretability, addressing missing data, and facilitating responsible deployment. Addressing these issues will be crucial for the effective translation of multimodal AI from research to routine biomedical practice.

Potential impact on clinical practices

Integrating AI into clinical bioimaging has the potential to enhance diagnostic accuracy and efficiency, and enable earlier and more precise disease detection, which is particularly impactful for conditions where timely intervention is critical. Specifically, by combining imaging data with patient-specific information, AI can support the development of tailored treatments and enable more precise prognostic predictions[173,174]. AI systems also support real-time decision making, which provides immediate insights during procedures and evaluations, and, through predictive modelling, can assist in long-term care planning[146].

Among recent innovations, AI-agents, systems that integrate AI models with biomedical databases, diagnostic tools, and analytical software, offer new opportunities for clinical bioimaging practice. These agents can streamline workflows by reducing the time clinicians spend on image analysis to allow greater focus on patient care. In particular, agentic MLLMs enable these agents to move beyond passive image interpretation. By leveraging their ability to autonomously select relevant imaging modalities, retrieve and synthesise biomedical information, perform complex analyses, and generate structured clinical reports, agentic MLLMs help unite analytical and operational tasks within a single system[175]. By combining these capabilities, AI-agents have the potential to improve consistency, efficiency, and the potential for more personalised and timely clinical decision-making.

To fully realise the potential of AI in clinical bioimaging, future developments must focus on creating AI models that are not only interpretable, but also capable of providing real-time explanations. These models should also be trained on diverse datasets that account for the variability in imaging conditions. At the same time, standardising imaging acquisition methods is essential to improve consistency. Addressing these barriers will help build trust in AI systems, enabling their large-scale adoption in clinics. With these improvements, AI-agents integrated with biomedical databases and clinical decision support tools will enhance accuracy, efficiency, and personalised

care, ultimately transforming healthcare by streamlining workflows, enabling earlier disease detection, and supporting long-term care planning.

Conclusion

AI techniques have demonstrated outstanding improvement in biomedical image analysis. While each biomedical imaging modality has unique characteristics, combining multiple imaging techniques via multimodal AI holds promise to improve patient outcomes and advance personalised medicine.

In this review, we discussed several AI techniques to integrate multimodal biomedical imaging, the pros and cons of each integration strategy and highlighted the emerging AI techniques (e.g., MLLMs and foundation models) that can further improve biomedical imaging integration. However, integrative analysis of multimodal biomedical imaging poses several challenges. Hence, we explored the key challenges in multimodal biomedical image analysis, including the availability of good-quality data, multimodal model interpretation and ethical considerations. Finally, we discussed emerging techniques and future directions for improving multimodal biomedical analysis.

In conclusion, multimodal AI holds huge potential in clinical translation due to its capability of providing real-time explanations. Moreover, the application of large-scale multimodal AI in the clinic can overcome data diversity challenges and provide a global standard for biomedical imaging analysis. However, further work is needed to effectively integrate biomedical imaging with other omic data, ultimately transforming healthcare by streamlining workflows, enabling earlier disease detection, and supporting long-term care planning.

Data availability

This review is based on previously published studies. All datasets mentioned are available in the public domain as indicated in the citations throughout the manuscript. No datasets were generated or analysed as part of the study.

Received: 10 February 2025; Accepted: 27 January 2026;
Published online: 13 February 2026

References

1. Imaging more of the body. *Nat. Biomed. Eng.* **6**, 495–496 (2022).
2. Bouma, B. E. et al. Optical coherence tomography. *Nat. Rev. Methods Prim.* **2**, 79 (2022).
3. Beaton, L., Bandula, S., Gaze, M. N. & Sharma, R. A. How rapid advances in imaging are defining the future of precision radiation oncology. *Br. J. Cancer* **120**, 779–790 (2019).
4. Khalifa, M. & Albadawy, M. AI in diagnostic imaging: revolutionising accuracy and efficiency. *Comp. Methods Program. Biomed. Update* **5**, 100146 (2024).
5. Park, J. et al. Clinical translation of photoacoustic imaging. *Nat. Rev. Bioeng.* **16**, 4353–4379 (2024).
6. Zhao, C., Guo, L., Dong, J. & Cai, Z. Mass spectrometry imaging-based multi-modal technique: next-generation of biochemical analysis strategy. *Innovation* **2**, 100151 (2021).
7. Duan, J., Xiong, J., Li, Y. & Ding, W. Deep learning based multimodal biomedical data fusion: an overview and comparative review. *Inf. Fusion* **112**, 102536 (2024).
8. Ma, J. et al. Segment anything in medical images. *Nat. Commun.* **15**, 654 (2024).
9. Li, M., Jiang, Y., Zhang, Y. & Zhu, H. Medical image analysis using deep learning algorithms. *Front. Public Health* **11**, 1273253 (2023).
10. Li, X., Zhang, L., Yang, J. & Teng, F. Role of artificial intelligence in medical image analysis: a review of current trends and future directions. *J. Med. Biol. Eng.* **44**, 231–243 (2024).
11. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
12. Zhang, K. et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nat. Med.* **30**, 3129–3141 (2024).
13. He, Y. et al. Foundation model for advancing healthcare: challenges, opportunities and future directions. *IEEE Rev. Biomed. Eng.* **18**, 172–191 (2024).
14. Ryu, J. S., Kang, H., Chu, Y. & Yang, S. Vision-language foundation models for medical imaging: a review of current practices and innovations. *Biomed. Eng. Lett.* **15**, 809–830 (2025).
15. Haq, I. U. et al. Advancements in medical radiology through multimodal machine learning: a comprehensive overview. *Bioengineering* **12**, 477 (2025).
16. Bahadir, C. D. et al. Artificial intelligence applications in histopathology. *Nat. Rev. Electr. Eng.* **1**, 93–108 (2024).
17. Guarrasi, V. et al. A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. *Image Vis. Comput.* **158**, 105509 (2025).
18. Warner, E. et al. Multimodal machine learning in image-based and clinical biomedicine: survey and prospects. *Int. J. Comp. Vis.* **132**, 1–17 (2024).
19. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nat. Med.* **28**, 1773–1784 (2022).
20. Rong, J. & Liu, Y. Advances in medical imaging techniques. *BMC Methods* **1**, 10 (2024).
21. Chen, J. & Chen, J. Multimodal image feature fusion for improving medical ultrasound image segmentation. *Biomed. Signal Process. Control* **89**, 105705 (2024).
22. Ottestad, A. L. et al. Associations between detectable circulating tumor dna and tumor glucose uptake measured by 18f-fdg pet/ct in early-stage non-small cell lung cancer. *BMC Cancer* **23**, 646 (2023).
23. Illimoottil, M. & Ginat, D. Recent advances in deep learning and medical imaging for head and neck cancer treatment: MRI, CT, and PET scans. *Cancers* **15**, 3267 (2023).
24. Kaniyala Melanthota, S. et al. Types of spectroscopy and microscopy techniques for cancer diagnosis: a review. *Lasers Med. Sci.* **37**, 3067–3084 (2022).
25. Zhang, H. et al. Mass spectrometry imaging for spatially resolved multi-omics molecular mapping. *npj Imaging* **2**, 20 (2024).
26. Doherty, T. et al. Feature fusion of Raman chemical imaging and digital histopathology using machine learning for prostate cancer detection. *Analyst* **146**, 4195–4211 (2021).
27. Nunes, J. B. et al. Integration of mass cytometry and mass spectrometry imaging for spatially resolved single-cell metabolic profiling. *Nat. Methods* **21**, 1796–1800 (2024).
28. Wang, D. et al. Automated machine-learning framework integrating histopathological and radiological information for predicting IDH1 mutation status in glioma. *Front. Bioinforma.* **1**, 718697 (2021).
29. Fitzgerald, S. et al. Multimodal raman spectroscopy and optical coherence tomography for biomedical analysis. *J. Biophotonics* **16**, e202200231 (2023).
30. Liu, Y., Valji, K., Monsky, W., Zheng, C. & Yang, X. Optical imaging guidance in oncologic surgery and interventional oncology. *Pharmacol. Res.* **212**, 107612 (2025).
31. Lin, L. & Wang, L. V. The emerging role of photoacoustic imaging in clinical oncology. *Nat. Rev. Clin. Oncol.* **19**, 365–384 (2022).
32. Smith, B., Hermsen, M., Lesser, E., Ravichandar, D. & Kremers, W. Developing image analysis pipelines of whole-slide images: pre-and post-processing. *J. Clin. Transl. Sci.* **5**, e38 (2021).
33. Cammarasana, S., Nicolardi, P. & Patanè, G. Real-time denoising of ultrasound images based on deep learning. *Med. Biol. Eng. Comput.* **60**, 2229–2244 (2022).
34. Chen, H. et al. Low-dose ct with a residual encoder-decoder convolutional neural network. *IEEE Trans. Med. Imaging* **36**, 2524–2535 (2017).
35. Gao, Q. & Shan, H. Cocodiff: a contextual conditional diffusion model for low-dose CT image denoising. In *Developments in X-Ray Tomography XIV*, **12242**, 92–98 (SPIE, 2022).

36. Lin, G. et al. Multimodal feature-guided diffusion model for low-count PET image denoising. *Med. Phys.* **52**, 4403–4415 (2025).
37. Parvaiz, A. et al. Vision transformers in medical computer vision-a contemplative retrospection. *Eng. Appl. Artif. Intell.* **122**, 106126 (2023).
38. Singh, A. R. et al. An enhanced denoising system for mammogram images using deep transformer model with fusion of local and global features. *Sci. Rep.* **15**, 6562 (2025).
39. Song, T.-A., Yang, F. & Dutta, J. Noise2void: unsupervised denoising of pet images. *Phys. Med. Biol.* **66**, 214002 (2021).
40. Shen, Y. et al. Pads-net: Gan-based radiomics using multi-task network of denoising and segmentation for ultrasonic diagnosis of parkinson disease. *Computerized Med. Imaging Graph.* **120**, 102490 (2025).
41. Daras, G. et al. A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083* (2024).
42. Chung, H., Kim, J. & Ye, J. C. Diffusion models for inverse problems. arXiv:2508.01975 (2025).
43. Kim, S. et al. Unsupervised learning for motion correction and assessment in brain magnetic resonance imaging using severity-based regularized cycle consistency. *Eng. Appl. Artif. Intell.* **142**, 109978 (2025).
44. Guo, K. et al. Medgan: An adaptive gan approach for medical image generation. *Computers Biol. Med.* **163**, 107119 (2023).
45. Lim, A., Lo, J., Wagner, M. W., Ertl-Wagner, B. & Sussman, D. Motion artifact correction in fetal mri based on a generative adversarial network method. *Biomed. Signal Process. Control* **81**, 104484 (2023).
46. Chen, Z. et al. Portable head ct motion artifact correction via diffusion-based generative model. *Computerized Med. Imaging Graph.* **119**, 102478 (2025).
47. Islam, T., Hafiz, M. S., Jim, J. R., Kabir, M. M. & Mridha, M. A systematic review of deep learning data augmentation in medical imaging: recent advances and future research directions. *Healthcare Anal.* **5**, 100340 (2024).
48. Dayarathna, S. et al. Deep learning based synthesis of MRI, CT and PET: review and analysis. *Med. image Anal.* **92**, 103046 (2024).
49. Nag, A. et al. Tumorganet: a transfer learning and generative adversarial network-based data augmentation model for brain tumor classification. *IEEE Access* **12**, 103060–103081 (2024).
50. Guan, Q. et al. Medical image augmentation for lesion detection using a texture-constrained multichannel progressive gan. *Computers Biol. Med.* **145**, 105444 (2022).
51. Pombo, G. et al. Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3d deep generative models. *Med. Image Anal.* **84**, 102723 (2023).
52. Zhang, Y. et al. Unified multi-modal image synthesis for missing modality imputation. *IEEE Trans. Med. Imaging* **44**, 4–18 (2024).
53. Saad, M. M., O'Reilly, R. & Rehmani, M. H. A survey on training challenges in generative adversarial networks for biomedical image analysis. *Artif. Intell. Rev.* **57**, 19 (2024).
54. Han, C. et al. Combining noise-to-image and image-to-image GANS: brain MR image augmentation for tumor detection. *IEEE Access* **7**, 156966–156977 (2019).
55. Liu, Q., Zhou, T., Cheng, C., Ma, J. & Hoque Tania, M. Hybrid generative adversarial network based on frequency and spatial domain for histopathological image synthesis. *BMC Bioinforma.* **26**, 29 (2025).
56. Qiang, N. et al. Modeling and augmenting of FMRI data using deep recurrent variational auto-encoder. *J. neural Eng.* **18**, 0460b6 (2021).
57. Huo, J. et al. Brain lesion synthesis via progressive adversarial variational auto-encoder. In *International Workshop on Simulation and Synthesis in Medical Imaging*, 101–111 (Springer, 2022).
58. Wei, R. & Mahmood, A. Recent advances in variational autoencoders with representation learning for biomedical informatics: a survey. *IEEE Access* **9**, 4939–4956 (2020).
59. Neloy, A. A. & Turgeon, M. A comprehensive study of auto-encoders for anomaly detection: efficiency and trade-offs. *Mach. Learn. Appl.* 17, 100572 (2024).
60. Kazerouni, A. et al. Diffusion models in medical imaging: a comprehensive survey. *Med. Image Anal.* **88**, 102846 (2023).
61. Yang, Z. et al. A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images. *Nat. Commun.* **16**, 2366 (2025).
62. Wu, C. et al. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nat. Commun.* **16**, 7866 (2025).
63. Van Griethuysen, J. J. et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **77**, e104–e107 (2017).
64. Singh, G. et al. Radiomics and radiogenomics in gliomas: a contemporary update. *Br. J. cancer* **125**, 641–657 (2021).
65. Boehm, K. M. et al. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat. Cancer* **3**, 723–733 (2022).
66. Zhang, X. et al. Deep learning with radiomics for disease diagnosis and treatment: challenges and potential. *Front. Oncol.* **12**, 773840 (2022).
67. Kalamkar, S. et al. Multimodal image fusion: a systematic review. *Decis. Anal. J.* 9, 100327 (2023).
68. Azam, M. A. et al. A review on multimodal medical image fusion: compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Computers Biol. Med.* **144**, 105253 (2022).
69. Huang, B., Yang, F., Yin, M., Mo, X. & Zhong, C. A review of multimodal medical image fusion techniques. *Comput. Math. methods Med.* **2020**, 8279342 (2020).
70. Liu, J., Dian, R., Li, S. & Liu, H. Sgfusion: a saliency guided deep-learning framework for pixel-level image fusion. *Inf. Fusion* **91**, 205–214 (2023).
71. Zhou, J. et al. A fusion algorithm based on composite decomposition for PET and MRI medical images. *Biomed. Signal Process. Control* **76**, 103717 (2022).
72. Li, X., Zhou, F., Tan, H., Zhang, W. & Zhao, C. Multimodal medical image fusion based on joint bilateral filter and local gradient energy. *Inf. Sci.* **569**, 302–325 (2021).
73. Odusami, M., Maskeliūnas, R. & Damaševičius, R. Pixel-level fusion approach with vision transformer for early detection of alzheimer's disease. *Electronics* **12**, 1218 (2023).
74. Bischof, J. et al. Multimodal bioimaging across disciplines and scales: challenges, opportunities and breaking down barriers. *npj Imaging* **2**, 5 (2024).
75. Yuan, Y. & Cheng, Y. Medical image segmentation with unet-based multi-scale context fusion. *Sci. Rep.* **14**, 15687 (2024).
76. Ma, J., Xu, H., Jiang, J., Mei, X. & Zhang, X.-P. DDCGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans. Image Process.* **29**, 4980–4995 (2020).
77. Tang, W. & He, F. Fatfusion: A functional–anatomical transformer for medical image fusion. *Inf. Process. Manag.* **61**, 103687 (2024).
78. Tang, W., He, F., Liu, Y. & Duan, Y. Matr: multimodal medical image fusion via multiscale adaptive transformer. *IEEE Trans. Image Process.* **31**, 5134–5149 (2022).
79. Bierbrier, J., Gueziri, H.-E. & Collins, D. L. Estimating medical image registration error and confidence: a taxonomy and scoping review. *Med. Image Anal.* **81**, 102531 (2022).
80. Wang, Z. & Hong, Y. A2fseg: Adaptive multi-modal fusion network for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 673–681 (Springer, 2023).

81. Bi, Y., Abrol, A., Fu, Z. & Calhoun, V. D. A multimodal vision transformer for interpretable fusion of functional and structural neuroimaging data. *Hum. Brain Mapp.* **45**, e26783 (2024).
82. Song, H. et al. Multimodal separation and cross fusion network based on Raman spectroscopy and FTIR spectroscopy for diagnosis of thyroid malignant tumor metastasis. *Sci. Rep.* **14**, 29125 (2024).
83. Chen, J. et al. A deep learning-based multimodal medical imaging model for breast cancer screening. *Sci. Rep.* **15**, 14696 (2025).
84. Thakoor, K. A. et al. A multimodal deep learning system to distinguish late stages of amd and to compare expert vs. AI ocular biomarkers. *Sci. Rep.* **12**, 2585 (2022).
85. Dasanayaka, S., Shantha, V., Silva, S., Meedeniya, D. & Ambegoda, T. Interpretable machine learning for brain tumour analysis using mri and whole slide images. *Softw. Impacts* **13**, 100340 (2022).
86. Zhang, T. et al. Predicting breast cancer types on and beyond molecular level in a multi-modal fashion. *npj Breast Cancer* **9**, 16 (2023).
87. Wu, J. et al. Multimodal microscopic imaging with deep learning for highly effective diagnosis of breast cancer. *Opt. Lasers Eng.* **168**, 107667 (2023).
88. Zhou, C. et al. Multimodal data integration for predicting progression risk in castration-resistant prostate cancer using deep learning: a multicenter retrospective study. *Front. Oncol.* **14**, 1287995 (2024).
89. Morais, M., Calisto, F. M., Santiago, C., Aleluia, C. & Nascimento, J. C. Classification of breast cancer in mri with multimodal fusion. In *2023 IEEE 20th international symposium on biomedical imaging (ISBI)*, 1–4 (IEEE, 2023).
90. Barai, A. et al. A late fusion deep cnn model for the classification of brain tumors from multi-parametric mri images. In *2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM)*, 1–6 (IEEE, 2023).
91. Zhang, S. & Metaxas, D. On the challenges and perspectives of foundation models for medical image analysis. *Med. Image Anal.* **91**, 102996 (2024).
92. Li, Z., Jiang, Y., Lu, M., Li, R. & Xia, Y. Survival prediction via hierarchical multimodal co-attention transformer: a computational histology-radiology solution. *IEEE Trans. Med. Imaging* **42**, 2678–2689 (2023).
93. Lin, C., Chen, Y., Feng, S. & Huang, M. A multibranch and multiscale neural network based on semantic perception for multimodal medical image fusion. *Sci. Rep.* **14**, 17609 (2024).
94. Liu, Y. et al. Brain tumor segmentation in multimodal mri via pixel-level and feature-level image fusion. *Front. Neurosci.* **16**, 1000587 (2022).
95. Zhang, X. et al. Development of a large-scale medical visual question-answering dataset. *Commun. Med.* **4**, 277 (2024).
96. Sheth, D. & Giger, M. L. Artificial intelligence in the interpretation of breast cancer on mri. *J. Magn. Reson. Imaging* **51**, 1310–1324 (2020).
97. Senthilkumar, K., Kalanandhini, G., Aravind, A., Vijayalakshmi, G. & Gayathri, J. Image fusion based on dtdwt to improve segmentation accuracy in tumour detection. In *AIP Conference Proceedings*, vol. 2393 (AIP Publishing, 2022).
98. Tang, X. et al. Preoperative prediction of axillary lymph node metastasis in patients with breast cancer through multimodal deep learning based on ultrasound and magnetic resonance imaging images. *Acad. Radiol.* **9**, 100327 (2024).
99. Oberije, C. J. et al. Assessing artificial intelligence in breast screening with stratified results on 306 839 mammograms across geographic regions, age, breast density and ethnicity: a retrospective investigation evaluating screening (aries) study. *BMJ Health Care Inform.* **32**, e101318 (2025).
100. Zhuang, X. et al. Cardiac segmentation on late gadolinium enhancement MRI: a benchmark study from multi-sequence cardiac MR segmentation challenge. *Med. Image Anal.* **81**, 102528 (2022).
101. Chen, J., Li, H., Zhang, J. & Menze, B. Adversarial convolutional networks with weak domain-transfer for multi-sequence cardiac mr images segmentation. In *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges: 10th International Workshop, STACOM 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Revised Selected Papers 10*, 317–325 (Springer, 2020).
102. Hatamizadeh, A. et al. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, 272–284 (Springer, 2021).
103. Shamshad, F. et al. Transformers in medical imaging: a survey. *Med. Image Anal.* **88**, 102802 (2023).
104. Boehm, K. M., Khosravi, P., Vanguri, R., Gao, J. & Shah, S. P. Harnessing multimodal data integration to advance precision oncology. *Nat. Rev. Cancer* **22**, 114–126 (2022).
105. Xia, Y. et al. CT-based multimodal deep learning for non-invasive overall survival prediction in advanced hepatocellular carcinoma patients treated with immunotherapy. *Insights into Imaging* **15**, 214 (2024).
106. Qi, Y. et al. Applications of Raman spectroscopy in clinical medicine. *Food Front.* **5**, 392–419 (2024).
107. Zhou, X., Chen, C., Zuo, E., Chen, C. & Lv, X. Cross branch co-attention network multimodal models based on raman and ftir spectroscopy for diagnosis of multiple selected cancers. *Appl. Soft Comput.* **166**, 112204 (2024).
108. Hsu, W.-W. et al. A weakly supervised deep learning-based method for glioma subtype classification using WSI and MPMRIS. *Sci. Rep.* **12**, 6111 (2022).
109. Lin, D. et al. Autofluorescence and white light imaging-guided endoscopic raman and diffuse reflectance spectroscopy for in vivo nasopharyngeal cancer detection. *J. Biophotonics* **11**, e201700251 (2018).
110. Lin, J., Li, Y., Lin, X. & Che, C. Decision-level data fusion based on laser-induced breakdown and raman spectroscopy: a study of bimodal spectroscopy for diagnosis of lung cancer at different stages. *Talanta* **275**, 126194 (2024).
111. Janßen, C. et al. Multimodal lung cancer subtyping using deep learning neural networks on whole slide tissue images and maldi msi. *Cancers* **14**, 6181 (2022).
112. Wu, C. et al. Multimodal fusion with hyperspectral images and rgb information for all-optical informed virtual staining. *Opt. Express* **33**, 19820–19836 (2025).
113. Chen, Z. et al. Multimodal optoacoustic imaging: methods and contrast materials. *Chem. Soc. Rev.* **53**, 6068–6099 (2024).
114. Kang, E. Y.-C. et al. A multimodal imaging–based deep learning model for detecting treatment-requiring retinal vascular diseases: model development and validation study. *JMIR Med. Inform.* **9**, e28868 (2021).
115. Chua, J. et al. Utilizing deep learning to predict alzheimer's disease and mild cognitive impairment with optical coherence tomography. *Alzheimer's. Dement. Diagn. Assess. Dis. Monit.* **17**, e70041 (2025).
116. Zhao, T. et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nat. Methods* **22**, 166–176 (2025).
117. Doan, L. M. T., Verma, S., Eftekhari, N., Angione, C. & Occhipinti, A. From bulk to single-cell and spatial data: an AI framework to characterise breast cancer metabolic dysregulations across modalities. *Computers Biol. Med.* **198**, 111195 (2025).
118. Verma, S. et al. Cross-attention enables deep learning on limited omics-imaging-clinical data of 130 lung cancer patients. *Cell Rep. Methods* **4**, 100817 (2024).
119. Qoku, A., Katsaouni, N., Flinner, N., Buettner, F. & Schulz, M. H. Multimodal analysis methods in predictive biomedicine. *Comput. Struct. Biotechnol. J.* **21**, 5829–5838 (2023).

120. Wu, J. et al. Integrating transcriptomics, genomics, and imaging in alzheimer's disease: a federated model. *Front. Radiol.* **1**, 777030 (2022).
121. Ning, C. et al. Genome-wide association analysis of left ventricular imaging-derived phenotypes identifies 72 risk loci and yields genetic insights into hypertrophic cardiomyopathy. *Nat. Commun.* **14**, 7900 (2023).
122. Gomes, B. et al. Genetic architecture of cardiac dynamic flow volumes. *Nat. Genet.* **56**, 245–257 (2024).
123. Yu, Y. et al. MRI-based multimodal AI model enables prediction of recurrence risk and adjuvant therapy in breast cancer. *Pharmacol. Res.* **216**, 107765 (2025).
124. Zhang, S. et al. Tumor initiation and early tumorigenesis: molecular mechanisms and interventional targets. *Signal Transduct. Target. Ther.* **9**, 149 (2024).
125. Lu, M. Y. et al. A multimodal generative AI copilot for human pathology. *Nature* **634**, 466–473 (2024).
126. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
127. Lu, M. Y. et al. A visual-language foundation model for computational pathology. *Nat. Med.* **30**, 863–874 (2024).
128. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
129. Rao, V. M. et al. Multimodal generative AI for medical image interpretation. *Nature* **639**, 888–896 (2025).
130. Pérez-García, F. et al. Exploring scalable medical image encoders beyond text supervision. *Nat. Mach. Intell.* **7**, 119–130 (2025).
131. Lee, J.-O., Zhou, H.-Y., Berzin, T. M., Sodickson, D. K. & Rajpurkar, P. Multimodal generative ai for interpreting 3d medical images and videos. *npj Digit. Med.* **8**, 1–8 (2025).
132. Tu, T. et al. Towards generalist biomedical AI, 2023. *arXiv: 2307.14334* (2023).
133. Wang, L. et al. Interpretable Bilingual Multimodal Large Language Model for Diverse Biomedical Tasks. In *International Conference on Learning Representations*. (eds. Yue, Y., Garg, A., Peng, N., Sha, F. & Yu, R.) Vol. 2025, 26054–26086 (2025).
134. Huang, X. et al. Towards a multimodal large language model with pixel-level insight for biomedicine. In *Proc. AAAI Conference on Artificial Intelligence*, **39**, 3779–3787 (AAAI, 2025).
135. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 84–90 (2012).
136. Li, Q. et al. Medical image classification with convolutional neural network. In *Proc. 13th International Conference on Control Automation Robotics & Vision (ICARCV)*, 844–848 (IEEE, 2014).
137. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241 (Springer, 2015).
138. Goodfellow, I. et al. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **27**, 2672–2680 (2014).
139. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. & Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265 (PMLR, 2015).
140. Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **278**, 563–577 (2016).
141. Gómez-de Mariscal, E. et al. Deep-learning-based segmentation of small extracellular vesicles in transmission electron microscopy images. *Sci. Rep.* **9**, 13211 (2019).
142. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 1–11 (2017).
143. Li, Y. et al. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. In *International Conference on Learning Representations*(2022).
144. AlSaad, R. et al. Multimodal large language models in health care: applications, challenges, and future outlook. *J. Med. Internet Res.* **26**, e59505 (2024).
145. Bradshaw, T. J. et al. Large language models and large multimodal models in medical imaging: A primer for physicians. *J. Nucl. Med.* **66**, 173–182 (2025).
146. Najjar, R. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics* **13**, 2760 (2023).
147. Schwabe, D., Becker, K., Seyferth, M., Klaß, A. & Schaeffter, T. The metric-framework for assessing data quality for trustworthy AI in medicine: a systematic review. *npj Digit. Med.* **7**, 203 (2024).
148. Karthik, A. et al. Comprehensive assessment of imaging quality of artificial intelligence-assisted compressed sensing-based MR images in routine clinical settings. *BMC Med. Imaging* **24**, 284 (2024).
149. Balachandar, N., Chang, K., Kalpathy-Cramer, J. & Rubin, D. L. Accounting for data variability in multi-institutional distributed deep learning for medical imaging. *J. Am. Med. Inform. Assoc.* **27**, 700–708 (2020).
150. Kondylakis, H. et al. Data infrastructures for ai in medical imaging: a report on the experiences of five eu projects. *Eur. Radiol. Exp.* **7**, 20 (2023).
151. Carrillo-Perez, F., Pizurica, M., Marchal, K. & Gevaert, O. Synthetic multimodal data modelling for data imputation. *Nat. Biomed. Eng.* **9**, 421–425 (2025).
152. Chen, B., Oliva, J. & Niethammer, M. A unified model for longitudinal multi-modal multi-view prediction with missingness. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 410–420 (Springer, 2024).
153. Qi, L. et al. A unified missing modality imputation model with inter-modality contrastive and consistent learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 44–53 (Springer, 2025).
154. Srinivasan, S. et al. Multimodal biomedical image segmentation using multi-dimensional u-convolutional neural network. *BMC Med. Imaging* **24**, 38 (2024).
155. Han, L. et al. All-in-one medical image-to-image translation. *Cell Rep. Methods* **5**, 101138 (2025).
156. Schäfer, R. et al. Overcoming data scarcity in biomedical imaging with a foundational multi-task model. *Nat. Comput. Sci.* **4**, 495–509 (2024).
157. Pahud de Mortanges, A. et al. Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. *npj Digit. Med.* **7**, 195 (2024).
158. Ghasemi, A., Hashtarkhani, S., Schwartz, D. L. & Shaban-Nejad, A. Explainable artificial intelligence in breast cancer detection and risk prediction: a systematic scoping review. *Cancer Innov.* **3**, e136 (2024).
159. Rotem, O. & Zaritsky, A. Visual interpretability of bioimaging deep learning models. *Nat. Methods* **21**, 1394–1397 (2024).
160. Sangnark, S., Rattanachaisit, P., Patcharatrakul, T. & Vateekul, P. Explainable multi-modal deep learning with cross-modal attention for diagnosis of dyssynergic defecation using abdominal x-ray images and symptom questionnaire. *IEEE Access* **12**, 78132–78147 (2024).
161. Occhipinti, A., Verma, S., Doan, L. M. T. & Angione, C. Mechanism-aware and multimodal ai: beyond model-agnostic interpretation. *Trends Cell Biol.* **34**, 85–89 (2024).
162. Shortliffe, E. H. & Chiang, M. F.*Biomedical Data: Their Acquisition, Storage, and Use*, 45–75 (Springer, 2021).
163. Nasir, S., Khan, R. A. & Bai, S. Ethical framework for harnessing the power of ai in healthcare and beyond. *IEEE Access* **12**, 31014–31035 (2024).
164. Berlinger, N. et al. Ethical framework for health care institutions responding to novel coronavirus SARS-COV-2 (COVID-19)

guidelines for institutional ethics services responding to COVID-19. *Hastings Cent.* **12**, 1–12 (2020).
165. Tejani, A. S., Ng, Y. S., Xi, Y. & Rayan, J. C. Understanding and mitigating bias in imaging artificial intelligence. *RadioGraphics* **44**, e230067 (2024).
166. Lekadir, K. et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* 388 (2025).
167. Sun, Z. et al. A scoping review on multimodal deep learning in biomedical images and texts. *J. Biomed. Inform.* **146**, 104482 (2023).
168. Walter, A. et al. Correlated multimodal imaging in life sciences: expanding the biomedical horizon. *Front. Phys.* **8**, 47 (2020).
169. Koetzier, L. R. et al. Generating synthetic data for medical imaging. *Radiology* **312**, e232471 (2024).
170. Pai, S. et al. Foundation model for cancer imaging biomarkers. *Nat. Mach. Intell.* **6**, 354–367 (2024).
171. Perez-Lopez, R., Ghaffari Laleh, N., Mahmood, F. & Kather, J. N. A guide to artificial intelligence for cancer researchers. *Nat. Rev. Cancer* **24**, 427–441 (2024).
172. Azad, B. et al. Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689* (2023).
173. Artsi, Y., Sorin, V., Glicksberg, B. S., Nadkarni, G. N. & Klang, E. Advancing clinical practice: the potential of multimodal technology in modern medicine. *J. Clin. Med.* **13**, 6246 (2024).
174. Kaczmarczyk, R., Wilhelm, T. I., Martin, R. & Roos, J. Evaluating multimodal AI in medical diagnostics. *npj Digit. Med.* **7**, 205 (2024).
175. Zhang, S., Dai, G., Huang, T. & Chen, J. Multimodal large language models for bioimage analysis. *Nat. Methods* **21**, 1390–1393 (2024).
176. Network, C. G. A. et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330 (2012).

Acknowledgements

CA would like to acknowledge funding from BBSRC (grant BB/Y01278X/1) and Horizon Europe (grant 101182278). AO acknowledges funding from EPSRC, United Kingdom, grant EP/Y001613/1, the British Society of Breast Radiologists (BSBR), and pilot funding from AIBIO-UK (BBSRC BB/Y006933/1). AO and CA also acknowledge support from COST action CA22103, A comprehensive network against brain cancer, Net4Brain, by COST (European Cooperation in Science and Technology).

Author contributions

L.M.T.D. and A.O. conceptualised the study. L.M.T.D wrote the original manuscript. K.S. contributed to discussions on preprocessing techniques, drafted the perspectives section and part of the challenges. Su.V. worked on the visualisation and contributed to preprocessing techniques and challenges. G.L. contributed to preprocessing techniques and challenges. A.M. worked on preprocessing techniques and compiled Table 1. Sn.V. contributed to Section 3. C.A. edited, reviewed and supervised the study. A.O. conceptualised, edited, reviewed, and supervised the study. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Inclusion & Ethics

This review did not involve the collection of new human or animal data, and therefore, ethical approval was not required. We aimed to include studies representing a broad range of populations, methodologies, and perspectives where available, to promote inclusivity and reduce bias. The review process was conducted in accordance with ethical standards for academic integrity.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44172-026-00602-x.

Correspondence and requests for materials should be addressed to Annalisa Occhipinti.

Peer review information *Communications Engineering* thanks Maria Zuluaga and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: [Anastasiia Vasylchenkova, Rosamund Daw]. A peer review file is available.