

Bridging Modalities with AI: A Review of AI Advances in Multimodal Biomedical Imaging

Corresponding Author: Professor Annalisa Occhipinti

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The rationale for multimodal integration is not clearly articulated; the paper lists imaging modalities but does not sufficiently explain how their complementary strengths justify integration.

Transitions between sections—particularly from single-modality descriptions to multimodal fusion techniques—are abrupt, leaving gaps in the logical flow of the narrative.

The discussion on preprocessing techniques lacks detailed technical depth; methods such as normalization, denoising, and feature extraction are mentioned superficially without critical comparison of traditional versus AI-driven approaches.

There is an absence of a rigorous comparative analysis among the proposed fusion strategies (pixel-level, feature-level, decision-level, and hierarchical fusion), especially regarding their performance under real-world conditions such as data misalignment and noise.

Figures, particularly Figures 1 and 2, are overly dense with information and include several abbreviations (e.g., SIFT, LBP, MedSAM) without adequate explanations, reducing their clarity.

The section on data augmentation, which discusses generative models (GANs, diffusion models, VAEs), does not adequately address the limitations or potential pitfalls associated with these methods in biomedical imaging applications.

The exploration of emerging techniques, such as foundation models and multimodal large language models, is too superficial given their potential impact, with little discussion on their inherent challenges or integration with existing imaging modalities.

While extensive literature is cited, the synthesis of previous works with the current review is mechanical, and the paper does not critically evaluate the quality or limitations of the cited studies.

The manuscript omits key literature on multimodal segmentation in cardiac MRI, such as "Cardiac Segmentation on Late Gadolinium Enhancement MRI: A Benchmark Study from Multi-Sequence Cardiac MR Segmentation Challenge," "Adversarial Convolutional Networks with Weak Domain-Transfer for Multi-Sequence Cardiac MR Images Segmentation," and "Deep Class-Specific Affinity-Guided Convolutional Network for Multimodal Unpaired Image Segmentation."

Reviewer #2

(Remarks to the Author)

Please see attached document for detailed comments. Thanks.

Reviewer #3

(Remarks to the Author)

This paper presents a thorough review of the biomedical image analysis pipeline when dealing with multimodal data. It

covers the different stages from pre-processing, data augmentation, fusion and, finally, the downstream tasks. The paper is well-written and easy to follow.

Main comments

- The title is misleading. It gives the impression the paper deals with foundation models, which is not the case. Please revise
- Multi-modal applications in biomedical imaging start to explore the usage of modalities beyond just images. New applications consider using text (reports) or electronic health records. This is a current trend [1] that should be presented in this review. Currently, it is just mentioned as a perspective
- The paper's core, which should be about the joint usage of multiple image modalities, is relatively little discussed. Most of the paper is devoted to discussing image modalities, which is great but perhaps better suited for a textbook, and pre-processing, which is common to many other applications and not specific to multi-modal image analysis. It would be good to see a more in-depth study of data fusion, the existing applications, and their technical limitations.

[1] Pérez-García, Fernando, et al. "Exploring scalable medical image encoders beyond text supervision." *Nature Machine Intelligence* (2025): 1-12

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

Thanks to the author for providing an extensive response to address my questions and concerns. I especially appreciate the author's efforts to revise Section 3 and 5, which have been improved a lot. Below is a follow-up questions with remaining concerns after reading the revised manuscript and point-to-point response:

1. The new categories look more reasonable, while it may still need more thinking about what is "data-driven imaging". It is not the image data collected from a physical scanner, while processed by some techniques to extract information. Is it a specific imaging modality or should it belong to an image processing procedure? It may be confusing for the reader since there is another section later to discuss about the image processing techniques.
2. After the revision, the structure and rationale of Section 2 still look kind of weird. It is not clear what is the highlight or key point in this section: to introduce different image processing tasks? Or different models? Or focus on multimodal images? It may be confusing to claim in this section: "we explore how several AI-driven techniques are applied in preprocessing the multi-modal biomedical images (Figure 1B and C) to improve image quality, handle missing modalities and identify regions of interest." Many of the works discussed in this section are only limited to only preprocessing but also downstream tasks. And how the missing modalities connected to other tasks are not clear either. This may also lead to confusion to understand Table 1. What is the key message or conclusion the paper expects the reader to extract from reading this long list of papers? It may be helpful to think further about it.
3. The author claims some key contributions that make this review distinctive from previous review papers. However, in this active field, there have been quite a few review articles which may have covered the several points claimed by the author to some extent. To just name a few as the list below. Thus, I am afraid that it may not be very convincing to claim these as unique contributions in the paper.

[1] Vision-language foundation models for medical imaging: a review of current practices and innovations, *Biomedical Engineering Letters*, 2025

[2] Multimodal Machine Learning in Image-Based and Clinical Biomedicine: Survey and Prospects, *IJCV*, 2024

[3] Advancements in Medical Radiology Through Multimodal Machine Learning: A Comprehensive Overview, *Bioengineering*, 2025

[4] A systematic review of intermediate fusion in multimodal deep learning for biomedical applications, *Image and Vision Computing*, 2025

[5] Foundation Model for Advancing Healthcare: Challenges, Opportunities and Future Directions, *IEEE Rev Biomed Eng.*, 2025

Reviewer #3

(Remarks to the Author)

The authors have undertaken a substantial amount of work to improve the manuscript. The new version is more solid and complete.

I recommend, however, revising Section 3.2 regarding the downstream tasks, as it remains weak. As this is a review paper, it should be motivated by an important amount of work in a domain (multi-modal biomedical imaging) that requires a survey summarizing the main achievements of the last couple of years (sometimes decades).

When reading Section 3.2, one gets the impression that the field has not reached a critical mass to "deserve" such a survey.

While some works are cited, these mainly cover multi-modal imaging applications that have been long established, such as multi-sequence MR in cardiac and brain imaging, or the combination of US and MR in cancer imaging. To make their point, the authors should consider expanding the section while providing concrete examples of emerging applications that go beyond classical ones, where novel multi-modal biomedical images are considered. A table summarizing these, would also be of great value.

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

Thank the author for the exhausted and detailed responses to answer my questions and address the concerns. Most of my concerns have been well addressed while there are a few follow-up points that may need further attention from the author.

- Q1: The author is encouraged to reconsider more about the category of "Data-driven imaging", since it is actually not a "physical imaging" modality but an image or data processing technique that can be essentially different from the other imaging categories. It may be more reasonable to be discussed in Section 2.

- Q3: Regarding the diffusion models for (biomedical) image restoration and reconstruction, it doesn't not mean the denoising process in DDPM. The author is encouraged to search the recent works in diffusion model for inverse problem solving. There have been many recent works in past few years studying this area with significant application in medical imaging, for example, just to list two review papers [A, B]. The description in the revised version may be misleading.

[A] Daras, Giannis, et al. "A survey on diffusion models for inverse problems." arXiv preprint arXiv:2410.00083 (2024).

[B] Chung, Hyungjin, et al. "Diffusion models for inverse problems." arXiv preprint arXiv:2508.01975 (2025).

Reviewer #3

(Remarks to the Author)

The authors have done a great work in improving the paper aiming to show that the different works that have been exploiting multimodal learning within the medical imaging field.

Overall, I think the paper is in good shape and can be considered for publication.

Nonetheless, I have one last relatively minor remark. In the challenges about data quality, the authors miss to mention the problem of missing modalities. This is a topic that has for long been disregarded, but that starts to gain traction as it is a critical problem that may hinder translation into practice, where a missing modality at inference can often occur. As pointers to some works that have started to look into this the authors may look into:

- https://doi.org/10.1007/978-3-031-72384-1_28

- https://doi.org/10.1007/978-3-031-84525-3_15

- https://doi.org/10.1007/978-3-031-72390-2_39

Please consider these (and perhaps others) in your discussion in section 4.1.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Reviewer #1

The rationale for multimodal integration is not clearly articulated; the paper lists imaging modalities but does not sufficiently explain how their complementary strengths justify integration.

We thank the reviewer for raising this point. We revised Section 1 highlighting how the unique strengths of each imaging modality can enhance biomedical imaging when combined, and clarify the rationale for multimodal integration, focusing on the complementary information of the modalities. This has been reported for all the imaging modalities described in the paper, as detailed below.

On page 3, we describe how multiple radiological imaging, such as ultrasound, MRI, CT scans, PET and SPECT can be integrated to provide more comprehensive diagnostic insights at the patient level, adding the following text:

“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) [14]. Additionally, PET and 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 [15]. 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 [16].”

We also elaborate on how the integration of other biomedical imaging, such as microscopic and spectroscopic imaging modalities, OCT and OI, can leverage the complementary information of each modality to have a deeper understanding of the physiological process at the patient, tissue, and cellular levels. This is reported on page 4, and also copied below:

“...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 can 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 [19]. 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) [20]. 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 [21]....

...The integration of Optical Coherence Tomography (OCT) and Optoacoustic Imaging (OI) with spectroscopic imaging (e.g., Raman and multi-spectral) complements the structural imaging with biochemical information of the tissue [22]. 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 [23, 24]. Specifically, the fusion of OI and MRI images 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 (TME) [24]...."

All these changes have been integrated into Section 1.1 in the updated manuscript.

Transitions between sections—particularly from single-modality descriptions to multimodal fusion techniques—are abrupt, leaving gaps in the logical flow of the narrative.

We appreciate the reviewer's valuable suggestion. To improve the logical flow of the manuscript and highlight the core of the review, we moved the details of single-modality descriptions to Box 1, allowing the main text to focus on multimodal integration. We also revised Section 1 to provide a clearer rationale for multimodal approaches, emphasising the complementary information from each modality to enhance the analysis and performance of downstream tasks. In Section 2, we improved the flow by elaborating on preprocessing techniques tailored to multimodal biomedical imaging, including specific examples such as image enhancement and handling missing modalities.

The discussion on preprocessing techniques lacks detailed technical depth; methods such as normalization, denoising, and feature extraction are mentioned superficially without critical comparison of traditional versus AI-driven approaches.

Thank you for raising this point. We have enhanced Section 2 by adding more in-depth technical details related to preprocessing techniques, focusing on critical evaluations between traditional and AI-driven approaches. We also placed more emphasis on technical details of other AI-based preprocessing techniques, such as augmentation, missing modality imputation and identification of regions of interest, in Sections 2.1, 2.2, and 2.3 (from pages 7 to 11).

Specifically, in each section, we added the traditional approach of each preprocessing technique and its limitations, and discussed how AI-based approaches overcome these limitations as follows:

Denoising and artefact correction: “...*Traditional image denoising methods (Figure 1B (ii)), including median filtering, Gaussian filtering, and total variation denoising, rely on mathematical assumptions and manually crafted priors. 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 [28]...*”

Regions of Interest (RoIs) identification and segmentation: “...*Classical methods for RoI detection, such as thresholding, edge detection, and region growing, have been widely used [41, 42]. For instance, thresholding classifies pixels based on intensity values, which works well for high-contrast images but often struggles with complex backgrounds or subtle intensity variations common in biomedical images [43]. While these traditional methods provide a coarse localisation of RoIs, AI-driven segmentation techniques offer a more precise, pixel-wise delineation of these regions, which is critical for accurate analysis in biomedical imaging....*”

Generative strategies for augmentation and missing modality imputation: “...*Traditional augmentation techniques, such as flipping, rotation, scaling and cropping, have been widely used to introduce statistical variability and increase diversity [49]. 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 reconstruct missing data or augment the data using available modalities, while incorporating clinical information [50]...*”

There is an absence of a rigorous comparative analysis among the proposed fusion strategies (pixel-level, feature-level, decision-level, and hierarchical fusion), especially regarding their performance under real-world conditions such as data misalignment and noise.

Following the reviewer’s suggestion, we have revised the text in Section 3.1 on pages 15 to 18, by adding a comparative discussion on the performance of different fusion strategies under real-world conditions (like data misalignment and noise).

We reported below the revised content (extracted from pages 15-18).

Pixel-level fusion: “... *Pixel-level fusion (also referred to as early fusion, Figure 2A(i)) combines different biomedical imaging modalities of the same sample region at a pixel-level to create a single and enhanced image [67]...*”

“...*For instance, combining pixel-level information from CT, MRI, and PET has been shown to enhance spatial and anatomical features of biomedical images [69-72]. 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 [73]. 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 [70, 74]...

...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...

...However, pixel-level fusion combines raw images and it can only be applied on 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, significantly 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 (See Section 2.1). 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 [78]..."

Feature-level fusion: "Feature-level fusion (also referred to as intermediate or joint fusion, Figure 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 [66]. 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 [66]. 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...

...Additionally, feature-level fusion also offers more flexibility in the regions of interest to be fused. 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 or structures within the body, enabling a more holistic understanding of the patient's condition... Histopathological imaging has been used to integrate 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 provides a more comprehensive perspective and enhances the clinical diagnosis [21]..."

"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, Figure 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 [66]. Moreover, by combining high-level decisions inferred from individual modalities, this fusion strategy has shown robust results in many applications and can handle the data misalignment and noise challenges between multiple imaging modalities. Decision-level fusion can integrate predictions from different biomedical imaging 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 subtype identification. ...Moreover, decision-level fusion has also significantly improved cancer patients' outcomes by integrating biomedical imaging from different biological views, such as radiomics from CT scans and histopathological features from WSI [25]....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, 89]...”

“...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 to train 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 (Figure 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 [66]...”

“...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... Despite its promising performance, combining multiple abstraction levels requires complex architectures, resulting in significantly higher computational costs compared to other fusion

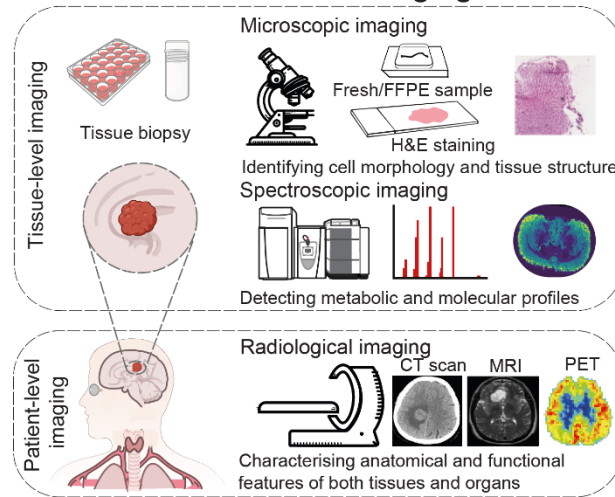
strategies. Although tailored approaches are applied to accommodate the differences in scale and resolution (e.g., Mixture of Experts (MoEs) where modality-specific expert networks are assigned to process different imaging modalities [93]), achieving effective optimisation and information fusion across heterogeneous modalities in hierarchical-level fusion remains a significant 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 significant improvements in model performance and flexibility, it still requires careful consideration of the trade-offs between model performance and complexity [66]..”

Figures, particularly Figures 1 and 2, are overly dense with information and include several abbreviations (e.g., SIFT, LBP, MedSAM) without adequate explanations, reducing their clarity.

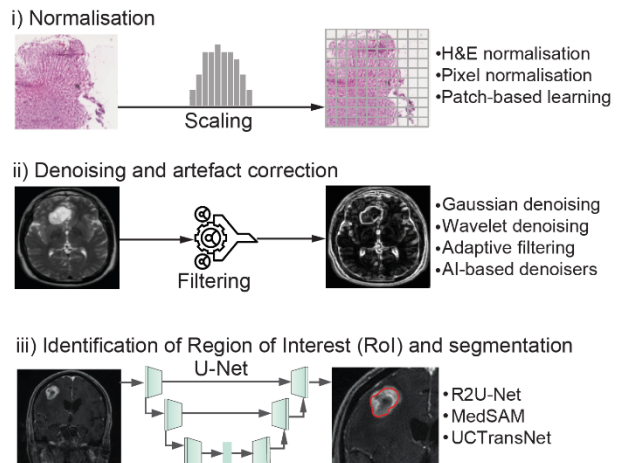
We appreciate the reviewer’s suggestion on improving the figures. We have revised both Figure 1 and Figure 2 and made them less cluttered. In particular, Figure 1A has been updated to demonstrate how several imaging modalities (microscopic, spectroscopic, and radiological images) can provide complementary information. Panels B and C have also been simplified to present the main biomedical image preprocessing categories. Similarly, we have also revised Figure 2 by clarifying abbreviations and making the sub-panels less dense.

The revised Figures 1 and 2 are reported below.

A. Multimodal biomedical imaging



B. Imaging preprocessing techniques



C. Generative models for data augmentation and missing modality imputation

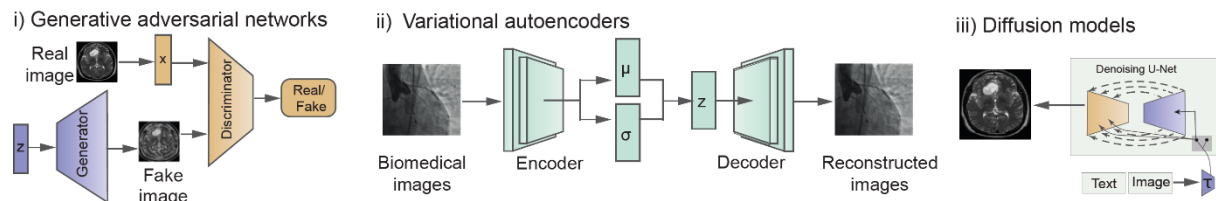
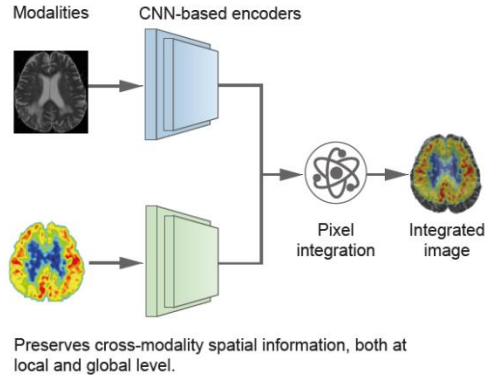


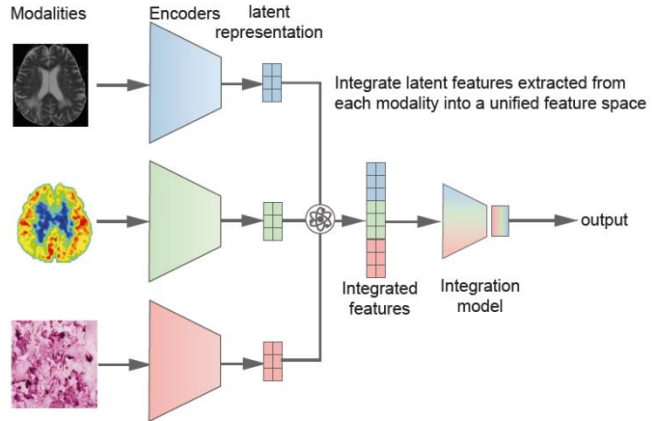
Figure 1: A) Multimodal biomedical imaging techniques, including microscopic, spectroscopic, and radiological methods, providing complementary biological insights into tissue function and structure, where brain tumour is considered as an example. Microscopic imaging captures the tissue architecture and cell morphology, while spectroscopic imaging detects metabolic and molecular profiles of the tissue. Similarly, radiological images (e.g., CT, MRI and PET) capture the patient-level image, characterising anatomical and functional features of both tissue and organ. B) Image preprocessing techniques including (i) normalisation, which allows for standardising the intensity values across images using H&E Normalisation. Gigapixel images are patched into smaller sizes and patch pixels are normalised; (ii) denoising and artefact correction, where noise and artefacts from the initial image are removed by applying filtering techniques (e.g., Gaussian- or Wavelet-denoising) and AI-based techniques, and (iii) identification of region of interest (RoI) and segmentation, which allows for the extraction of the region of interest from the starting image (R2U-Net: Recurrent Residual U-Net; MedSAM: Segment Anything in Medical Images; UCTransNet: Channel Transformer module in U-Net). These preprocessing techniques are usually adopted before training the multimodal AI model to enhance its performance. C) Strategies for image augmentation and handling missing modalities using generative models: (i) generative adversarial networks, which are usually applied for image reconstruction and cross-modality synthesis in medical imaging; (ii) variational autoencoders, which can generate synthetic biomedical images from Gaussian distribution, and (iii) diffusion models, which allow for robust image generation by iteratively refining a noisy image towards a clean image, guided by a learned noise model.

A. Multimodal fusion strategies

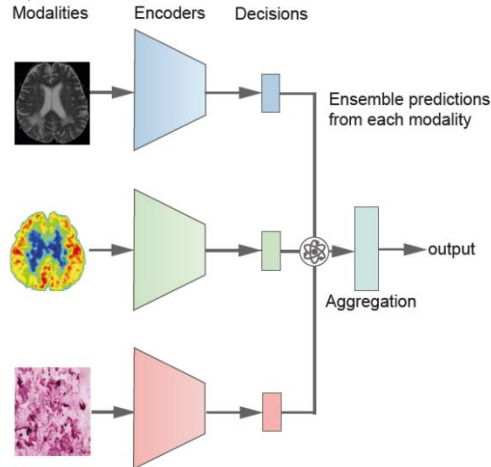
i) Pixel-level fusion



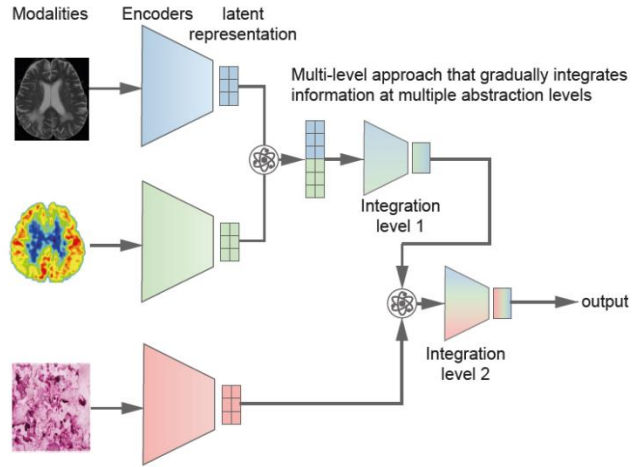
ii) Feature-level fusion



iii) Decision-level fusion

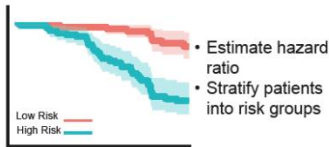


iv) Hierarchical-level fusion

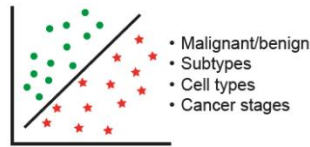


B. Downstream tasks and biological insights

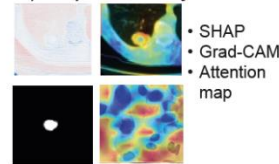
i) Risk prediction



ii) Classification tasks



iii) Explainability



iv) Tissue analysis

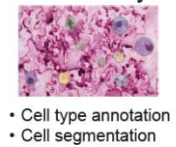


Figure 2: A) AI fusion strategies to combine and analyse multimodal biomedical images. (i) Pixel-level fusion: raw pixel intensity values from different imaging modalities (e.g., MRI and PET) are combined to generate superimposed images; **(ii) Feature-level fusion:** latent representations from each 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., SHAP, 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.

The section on data augmentation, which discusses generative models (GANs, diffusion models, VAEs), does not adequately address the limitations or potential pitfalls associated with these methods in biomedical imaging applications.

We thank the reviewer for this observation. In the revised manuscript, we have substantially modified Section 2 to more thoroughly address the limitations and potential pitfalls of generative models in biomedical imaging. Specifically, on pages 9 and 10, we have now elaborated on critical concerns such as mode collapse, non-convergence, and instability in GANs; the high computational burden and limited anatomical consistency in diffusion models; and issues like blurry reconstructions, posterior collapse, and limited interpretability in VAEs. We also discussed how these challenges might compromise clinical reliability and data efficiency, which are crucial in biomedical applications. Furthermore, we cited recent literature proposing mitigation strategies and emphasised that despite ongoing advancements, ensuring robust and trustworthy application of these models in clinical settings remains an open research challenge.

We have revised the corresponding section as follows.

“...GANs (Figure 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 [51–53]. 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 [54]. 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 [55, 56]. To mitigate these issues, recent works have explored architectural improvements, regularisation strategies, and hybrid model designs [55, 57]. 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 preprocessing for synthetic data augmentation (Figure 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 [58, 59]. Although some variants of VAEs have been explored for missing modality reconstruction, their use in this context is limited in favour of GANs and transformer models. In fact, VAEs pose challenges such as blurry outputs, weak cross-modal fidelity, and latent space entanglement [60]. Enhancements like autoregressive 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 [61]. ...”

“...Beyond GANs and VAEs, other deep-learning frameworks have also been developed to robustly address missing modality imputation. These include diffusion models (Figure 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 [62]. In tasks such as brain tumour segmentation, where acquiring a complete set of sequences (e.g., T1, T2, FLAIR, and contrast-enhanced MRI) is often unfeasible, models like M3AE leverage self-supervised learning to reconstruct missing modalities using the available ones [63,64]. Techniques such as modality dropout during training enable these models to generalise across varying modality combinations, while attention-based mechanisms improve feature fusion across present modalities [65].

The exploration of emerging techniques, such as foundation models and multimodal large language models, is too superficial given their potential impact, with little discussion on their inherent challenges or integration with existing imaging modalities.

We thank the reviewer for raising this important point. In the Perspective section of the revised manuscript (Section 5 on page 23) we have expanded the discussion on foundation models and multimodal large language models (MMLM), by providing a more in-depth analysis of their emerging role and challenges in multimodal bioimaging, specifically highlighting:

- The advantages of foundation models in transfer learning and generalisation across diverse and unlabelled datasets, which could resolve important issues in the bioimaging field (in Section 5.1, on page 24).

“... Foundation models and multimodal large language models (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 and mitigates the variability introduced by different imaging protocols and sources [129]...”

- How MLLMs integrate image and text data to provide richer contextual understanding and improved explainability, and enable zero-shot learning towards generalist AI models (in Section 5.3, on page 26).

“...MLLMs extend this potential by integrating text-based metadata with imaging data, enabling richer contextual understanding [93]. These models boost the capabilities of foundation models and remove the need for task-specific fine-tuning [130], although their application in biomedical imaging still faces significant 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 [131]. 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...”

- Critical challenges such as semantic misalignment between visual and textual data in bioimaging, difficulties in aligning heterogeneous imaging modalities (e.g., patch-based

WSIs versus radiology scans), interpretability concerns, and the high computational and privacy costs of MLLMs (in Section 3.1, on page 18 and Section 4.3, on pages 22 and 23).

“While attention mechanisms and image-text grounding can improve explainability in MLLMs [129], 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 [132, 133]. 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....”

These additions aim to present a more balanced and technically grounded discussion to reflect the potential and limitations of these technologies in bioimaging.

While extensive literature is cited, the synthesis of previous works with the current review is mechanical, and the paper does not critically evaluate the quality or limitations of the cited studies.

We thank the reviewer for this comment. To address this point, we have now added a critical evaluation of the cited studies. The changes related to this point are reported in Section 2 from page 7 to page 11 of the revised manuscript. Specifically, the evaluation of the cited studies related to AI-driven preprocessing techniques is reported below.

Biomedical image denoising and artefact correction (Section 2.1 on page 7): “...**Supervised deep learning, such as CNN-based models like RED-CNN [29], and perceptual frameworks using GANs, have shown strong performance in biomedical denoising and artefact correction [30, 31]. Furthermore, transformer-based models like DeepTFormer, TED-net and TransCT improve denoising by leveraging global context through attention mechanisms [32, 33]. However, these supervised methods require large 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 [34, 35]...For artefact correction, such as mitigating motion artefacts in MRI or enhancing low-dose CT scans, GANs 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 [36, 37]...**”

Regions of Interest (Rols) identification and segmentation (Section 2.2 on page 9):

“Classical methods for Rol detection, such as thresholding, edge detection, and region growing, have been widely used [41,42]...

...Deep learning-based methods have transformed segmentation with architectures like U-Net and its derivatives, including nnU-Net and R2U-Net, which utilise convolutional layers for multi-scale spatial information and hierarchical feature extraction, achieving superior performance in both 2D and 3D image segmentation [45]. For example, dual U-Net strategy has been effectively applied to cardiac magnetic resonance imaging to enhance segmentation results across multiple imaging sequences [46]. Transformer-based models, such as Swin UNETR, Tran-sUNet, and UCTransNet, 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 [47, 48]. These models are particularly effective for cerebral and liver MRI due to their ability to model complex anatomical structures...

However, despite their strengths, these models can struggle with scalability, cross-modality generalisation, and adaptability to varying clinical scenarios...

Foundation models like MedSAM further push the boundaries of multimodal segmentation 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 significant step towards universal biomedical segmentation solutions [8]...”

Generative strategies for augmentation and missing modality imputation (Section 2.3 on pages 9-11): “ ... By generating anatomically plausible synthetic images, GANs 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 [51–53]. 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 [54]. 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 [55, 56]. To mitigate these issues, recent works have explored architectural improvements, regularisation strategies, and hybrid model designs [55, 57]. Despite these advances, further validation is needed to ensure that generated images support consistent and interpretable outcomes in clinical workflows...

....In fact, VAEs pose challenges such as blurry outputs, weak cross-modal fidelity, and latent space entanglement [60]. Enhancements like autoregressive 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 [61]...Particularly, latent Diffusion Models (LDMs) can effectively capture long-range dependencies, enabling the generation of realistic MRI synthesis [62]. In tasks such as brain tumour segmentation, where acquiring a complete set of sequences (e.g., T1, T2, FLAIR, and contrast-enhanced MRI) is often unfeasible, models like M3AE leverage self-supervised

learning to reconstruct missing modalities using the available ones [63, 64]...Despite their promising performance in missing modality imputation, translating these AI-driven approaches remains a challenge due to lack of standardised imaging process, domain shifts and protocol variability."

The manuscript omits key literature on multimodal segmentation in cardiac MRI, such as "Cardiac Segmentation on Late Gadolinium Enhancement MRI: A Benchmark Study from Multi-Sequence Cardiac MR Segmentation Challenge." "Adversarial Convolutional Networks with Weak Domain-Transfer for Multi-Sequence Cardiac MR Images Segmentation," and "Deep Class-Specific Affinity-Guided Convolutional Network for Multimodal Unpaired Image Segmentation."

We thank the reviewer for mentioning additional relevant studies on multimodal segmentation in cardiac MRI. Segmentation is one of the downstream applications of multimodal applications that we discussed in Section 3. We have now incorporated the suggested studies into Section 3.2 in the revised manuscript, as below:

"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, enhance the accuracy of tumour and lesion detection, used for preoperative planning and prediction of metastasis. Integrating complementary information from multiple sequencing techniques in cardiac magnetic resonance (e.g., auxiliary and late gadolinium enhancement) has improved the segmentation results and reliability compared to one imaging modality only [46, 97]."

Reviewer #2

This review explores and discusses the recent artificial intelligence (AI) techniques applied in multimodal biomedical imaging, while also presenting key challenges and future directions. It outlines several integration strategies for combining diverse biomedical imaging data types and addresses challenges related to data quality, model interpretability, and ethical considerations. The manuscript is well-written and thoughtfully organized, featuring helpful illustrations—such as figures and tables—that effectively summarize relevant works and recent advances in the field of multimodal biomedical imaging.

While the paper presents interesting content and clear demonstrations, there are several areas where further clarification and refinement may enhance this review article.

- Section 1.1 provides a broad overview of various biomedical imaging modalities, including radiological imaging, microscopic imaging, histopathological imaging, etc. This comprehensive summary is nice. However, the categorization of imaging modalities could benefit from clearer distinctions, as some categories appear to overlap or exist at different hierarchical levels. For instance, ultrasound imaging is a subset of radiological imaging; microscopic imaging can include both histological and histopathological techniques; and modalities such as multiplexed or data-driven imaging may refer more towards the image processing and analytical approaches rather than to distinct imaging types. It may be worthwhile to consider reorganizing the categories to better reflect their functional or hierarchical relationships and improve conceptual clarity.

We thank the reviewer for this suggestion. In the revised manuscript, we have reorganised the biomedical imaging categories and subcategories in Section 1.1 and Box 1, avoiding any references to multiplex imaging as a single imaging modality, as it might overlap with other imaging types, such as microscopic and spectroscopic imaging. In the revised manuscript, we included four main categories of biomedical imaging modalities (as reported in Box 1 on page 4) and summarised below.

1. **Radiological imaging** includes X-rays, mammography, ultrasound, MRI, CT, PET and SPECT.
2. **Microscopic and spectroscopic imaging** include (but are not limited to) histological and histopathological imaging, Raman spectroscopy, and fluorescence imaging.
3. **Optical imaging**, including Optical Coherence Tomography (OCT) and Optoacoustic Imaging (OI), utilise light to detect anomalies within tissues.
4. **Data-driven imaging** techniques include radiomics, deriving tumour phenotypes (e.g., texture and shape) and radiogenomics, correlating radiomics features with genetic profiles.

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.

4. Data-driven imaging

Data-driven imaging techniques leverage AI-driven methods to extract meaningful patterns from complex biomedical images (e.g., radiological), enabling the detection of abnormalities and assisting in disease diagnosis, prognosis, and precision medicine.

Radiomics derives tumour information (e.g., tumour texture, shape, and intensity) from radiological images to characterise tumour phenotypes.

Radiogenomics correlate radiomics features with genomic data to uncover associations between imaging phenotypes and underlying molecular characteristics.

Furthermore, we acknowledge that data-driven imaging, such as radiomics, could refer to image processing and analytical techniques that, for example, extract tumour phenotype information from radiological imaging. To highlight the emerging role of data-driven imaging, we added more details in Box 1 and Section 1 (on page 4), as below.

“...Finally, data-driven imaging, including radiomics and radiogenomics (Box 1), represents a significant advancement in the biomedical field, driven by AI innovations to provide more insights into the disease. Radiomics derives tumour information (e.g., tumour texture, shape, and intensity) from radiological images, and radiogenomics aims to correlate 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 disease. With the support of AI, radiomics and radiogenomics are widely integrated with several biomedical imaging modalities (e.g., radiological and histopathological images) to enhance risk stratification, provide accurate diagnosis, and improve personalised cancer treatments [25,26]...”

Section 1 is now focused on multimodal biomedical imaging and how the complementary information of multiple biomedical imaging modalities could be combined to improve our understanding of biological processes at the organ, tissue, and cellular levels and improve diagnostic accuracy.

- Section 2 raises similar concerns regarding the logical structure of the discussion. From an image processing perspective, techniques such as normalization and denoising are typically components of a broader category of image enhancement, which may also include deblurring, super-resolution and etc., depending on the imaging modality and task requirements.

Additionally, while feature extraction is often associated with convolutional neural networks (CNNs) in modern contexts, it may not be limited in this scope. Traditional hand-crafted methods such as Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT) remain relevant and effective.

Furthermore, tasks like image segmentation are more accurately described as distinct image analysis objectives, similar to classification, detection, and localization. Clarifying these distinctions could help improve the logical flow of the section.

We thank the reviewer for this comment. We have restructured the discussion in Section 2 to better reflect the flow of the AI-driven preprocessing techniques in the multimodal biomedical imaging context. The three new subsections in Section 2 are reported below.

- Section 2.1: “Biomedical image denoising and artefact correction” discusses the traditional and AI-driven approaches (e.g., VAEs and diffusion models) to denoise

biomedical images and artefact correction.

- Section 2.2: “Regions of Interest (Rols) identification and segmentation” is focussed on preprocessing techniques to identify and segment Rols, which is a critical intermediate step to improve model performance.
- Section 2.3: “Generative strategies for augmentation and missing modality imputation” discusses the use of generative models, such as VAEs and GANs, in multimodal biomedical data augmentation and impute missing modality.

We also agree that feature extraction and segmentation may not be limited to the preprocessing scope. Indeed, feature extraction is an important intermediate step in the AI pipeline, particularly within feature-level integration for multimodal biomedical imaging. Following the reviewer’s suggestions, we moved the feature extraction discussion based on CNN-based models in the preprocessing step and expanded its discussion in the feature-level integration section (Section 3.1 on page 16), as reported below.

“...By leveraging extracted features rather than raw pixel data from each modality, feature-level fusion has demonstrated more robustness to handle misalignment and noise than pixel-level fusion [66]. 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 [79–81]...”

We also highlighted the role of image segmentation as part of image analysis objectives in identifying RoI, helping reduce computational complexity, and improving the performance and reliability of diagnostic prediction. These changes are reported in Section 2.2 on page 9 and copied below.

“...Segmentation addresses the challenge of isolating visually similar or overlapping components, such as organs, tumours, or cell nuclei, from surrounding tissues in complex biomedical images [44]. Deep learning-based methods have transformed segmentation with architectures like U-Net and its derivatives, including nnU-Net and R2U-Net, which utilise convolutional layers for multi-scale spatial information and hierarchical feature extraction, achieving superior performance in both 2D and 3D image segmentation [45]. For example, dual U-Net strategy has been effectively applied to cardiac magnetic resonance imaging to enhance segmentation results across multiple imaging sequences [46]. Transformer-based models, such as Swin UNETR, TransUNet, and UCTransNet, 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 [47,48]. These models are particularly effective for cerebral and liver MRI due to their ability to model complex anatomical structures. However, despite their strengths, they can struggle with scalability, cross-modality generalisation, and adaptability to varying clinical scenarios.

Foundation models like MedSAM further push the boundaries of multimodal segmentation 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 significant step towards universal biomedical segmentation solutions [8]...”

- In Section 2.1, the use of generative models such as GANs, VAEs, and diffusion models for data augmentation is mentioned. While generative models can indeed assist with data augmentation, it is worth noting that their effectiveness in improving downstream predictive performance remains an area of active research with existing debating on this topic. Therefore, it may be advisable to present a more balanced view by noting these potential issues and discussions, and avoid labeling them only as data augmentation approaches. Instead, these generative models have shown to be very useful in image reconstruction and restoration tasks in biomedical imaging applications, in the cutting-edge research.

We appreciate the reviewer’s comment and we agree that a more balanced perspective is important. We have revised Section 2 (from page 6 to page 11) to explicitly acknowledge that while generative models such as GANs, VAEs, and diffusion models offer powerful capabilities for data augmentation, their effectiveness in improving downstream predictive performance is still an open and actively debated research question. We also provided a more in-depth discussion of the use of these generative models and their limitations from a more balanced view.

The revised text from Section 2 is reported below.

Biomedical image denoising and artefact correction (Section 2.1):

“...Supervised deep learning, such as CNN-based models like RED-CNN [29], and perceptual frameworks using GANs, have shown strong performance in biomedical denoising and artefact correction [30, 31]. Furthermore, transformer-based models like DeepTFormer, TED-net and TransCT improve denoising by leveraging global context through attention mechanisms [32, 33]. However, these supervised methods require large 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 [34, 35]. Additionally, generative models like VAEs and Denoising Diffusion Probabilistic Models (DDPMs) 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. DDPMs (Figures 1C (ii)), 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. For artefact correction, such as mitigating motion artefacts in MRI or enhancing low-dose CT scans, GANs 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 [36, 37]. Similarly, diffusion-based models, such as the conditional diffusion model for portable head CT, iteratively refine images by modelling noise distributions, demonstrating significant improvements in motion artefact correction [38, 39]....

Generative strategies for augmentation and missing modality imputation (Section 2.3):

“...GANs (Figures 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, GANs 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 [51–53]. 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 [54]. 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 [55, 56]. To mitigate these issues, recent works have explored architectural improvements, regularisation strategies, and hybrid model designs [55, 57]. 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 preprocessing for synthetic data augmentation (Figures 1C (iii)). 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 [58, 59]. Although some variants of VAEs have been explored for missing modality reconstruction, their use in this context is limited in favour of GANs and transformer models. In fact, VAEs pose challenges such as blurry outputs, weak cross-modal fidelity, and latent space entanglement [60]. Enhancements like autoregressive 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 [61].

Beyond GANs and VAEs, other deep-learning frameworks have also been developed to robustly address missing modality imputation. These include diffusion models (Figures 1C (ii)), 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 [62]. In tasks such as brain tumour segmentation, where acquiring a complete set of sequences (e.g., T1, T2, FLAIR, and contrast-enhanced MRI) is often unfeasible, models like M3AE leverage self-supervised learning to reconstruct missing modalities using the available ones [63, 64]. Techniques such as modality dropout during training enable these models to generalise across varying modality combinations, while attention-based mechanisms improve feature fusion across present modalities [65]. Despite their promising performance in missing modality imputation, translating these AI-driven approaches remains a challenge due to lack of standardised imaging process, domain shifts

and protocol variability....”

- Section 3.1 presents an overview of multimodal image fusion strategies categorized as pixel-level, feature-level, decision-level, and hierarchical-level approaches. While this framework is informative, it is important to note that prior literature often refers to these strategies as early, intermediate, late, and joint fusion, respectively, as discussed in [1]. Connecting these terms to existing literature would strengthen the contextual grounding of the discussion.

We thank the reviewer for bringing this to our attention. We have now incorporated these connections into Section 3.1 in our manuscript, explicitly mapping early, intermediate (joint), and late fusion to pixel-level, feature-level, and decision-level strategies, respectively. The revised text from Section 3.1 is reported below.

“Pixel-level fusion (also referred to as early fusion, Figure 2A(i)) combines different biomedical imaging modalities of the same sample region at a pixel-level to create a single and enhanced image...

Feature-level fusion (also referred to as intermediate or joint fusion, Figure 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...

...in decision-level fusion (also referred to as late fusion, Figure 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 ...”

- Table 2 offers a helpful summary of different previous works based on various integration techniques. To further enhance the review, it may also be beneficial to include an additional table categorizing existing works based on specific tasks or anatomical sites. This could provide a clearer benchmark of how multimodal images, or different approaches address particular clinical or analytical challenges.

Thank you for this insightful suggestion. We have revised Table 2 to include an additional column that captures the anatomical site addressed in each study. This provides a more structured overview of how multimodal imaging techniques and integration approaches are applied to various clinical challenges across different anatomical regions. We believe this addition strengthens the comprehensiveness and practical applicability of the review.

The revised table is copied below.

Integration techniques	Imaging modalities	Downstream tasks	Anatomical sites	Ref
Pixel-level	MRI, PET	Preserving imaging details from both modalities to enhance clinical diagnosis	Brain	[69]
	MRI, PET	Enhancing image quality and visualisation (higher brightness and clearer edges) to improve the clinical diagnosis	Brain	[70]
	MRI (T1, T2), PET, SPECT	Preserving imaging details from both modalities to enhance clinical diagnosis	Brain	[71]
	MRI, PET	Avoiding blurring and reducing visible texture detail loss to improve the clinical diagnosis	Brain	[75]
	MRI, SPECT	Improve radiological image characterisation by combining anatomical soft-tissue structures with functional metabolic features in brain imaging	Brain	[76, 77]
	MRI, PET	Alzheimer’s disease detection	Brain	[72]
Feature-level	OCTA, OCT (structure), ultrasound (B-scan flow)	Age-related macular degeneration diagnosis	Eyes	[82]
	Raman, FTIR spectroscopy	Thyroid tumour diagnosis	Cervical region	[81]
	MRI and fMRI	Schizophrenia diagnosis	Brain	[80]
	MRI (T1, T1c, T2, FLAIR)	Brain tumour	Brain	[79]
	Raman, FTIR spectroscopy	Cancer diagnosis (e.g., lung cancer, glioma, and thyroid)	Lung, brain, neck	[101]
	WSI, MRI	Mutation status prediction	Brain	[21]
	Mammography, ultrasound	Breast cancer subtype identification	Breast	[84]
Decision-level	Radiomics (extracted from CT), WSI	Ovarian cancer risk stratification	Ovary	[25]
	Microscopic imaging (AFMI, BFMI, and OPMI)	Breast cancer diagnosis	Breast	[85]
	Radiomics (extracted from CT), WSI	Prostate cancer progression risk prediction	Prostate gland	[86]
	MRI, WSI	Brain tumour subtype identification	Brain	[83]
	MRI (T1, T2, DCE)	Breast cancer diagnosis	Breast	[87]
	MRI (T1, T1c, T2, FLAIR)	Brain tumour subtype identification	Brain	[88]
Hierarchical-level	MRI (T1, T1c, T2, FLAIR)	Brain tumour segmentation	Brain	[92]
	CT, WSI	Gastric cancer diagnosis	Stomach	[90]
	CT, MRI	Brain image quality enhancement	Brain	[91]

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. 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, MRI: Magnetic Resonance Imaging (T1, T1-weighted imaging with contrast (T1c), T2, Fluid-Attenuated Inversion Recovery (FLAIR), Dynamic Contrast-Enhanced (DCE) are different types of MRI sequences), 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.

- Regarding the discussion of multimodal large language models (MLLMs) as a prospective future direction, it is worth highlighting that recent studies have already begun to demonstrate their potential in biomedical imaging contexts. A more detailed review of these emerging works, possibly integrated into the main text rather than limited to a brief mention in the future directions, would add great value to make this review article more interesting and timely.

Thank you for this insightful suggestion. Following this recommendation, we have expanded the discussion on MLLMs to integrate a thorough discussion of recent advancements and practical demonstrations of MLLMs within biomedical imaging. Specifically, we now discuss:

- How MLLMs integrate image and text data to provide a richer contextual understanding and improved explainability and enable zero-shot learning towards generalist AI models (in Section 5.3, on page 26).
- Critical challenges such as semantic misalignment between visual and textual data in bioimaging, difficulties in aligning heterogeneous imaging modalities (e.g., patch-based WSIs vs. radiology scans), interpretability concerns, and the high computational and privacy costs of MLLMs (Section 4.3, on pages 22 and 23).
- Incorporating concrete, up-to-date examples of their use in biomedical imaging in Section 5.3 on page 26:
 - o Histopathology applications: We now cite Lu et al. (2024) and Chen et al. (2024), who demonstrate how MLLMs can link whole-slide image features with pathology reports to improve diagnostic classification.
 - o Radiology use cases: We discuss recent studies by Rao et al. (2025), Pérez et al. (2025), and Lee et al. (2025) that apply MLLMs for zero-shot recognition of radiological findings through image–text grounding.
 - o Multi-modality frameworks: We highlight approaches from Tu et al. (2023), Wang et al. (2024), and Huang et al. (2025) that propose large models to integrate multiple imaging modalities alongside text.
- Technical challenges and solutions: We add a balanced discussion of semantic alignment between visual tokens and language, interpretability limits of attention mechanisms, and the computational overhead of MLLMs, alongside emerging strategies such as pixel-aware grounding and lightweight architectures to address these issues (in Section 5.1, on page 24).

These additions strengthen the manuscript by providing timely and concrete examples that better illustrate the current state-of-the-art and potential of MLLMs in multimodal bioimaging, as reported below.

“... Foundation models and multimodal large language models (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 and mitigates the

variability introduced by different imaging protocols and sources [129].

MLLMs extend this potential by integrating text-based metadata with imaging data, enabling richer contextual understanding [93]. These models boost the capabilities of foundation models and remove the need for task-specific fine-tuning [130], although their application in biomedical imaging still faces significant 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 [131]. 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 [129], 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 [132, 133]. 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....”

- Similarly, Section 5.3, which discusses the integration of imaging data with other data modalities, touches on a highly active area of current research. Expanding this section to include a more thorough survey of recent advancements would likely resonate with readers interested in this growing subfield.

We thank the reviewer for their suggestion. We have extended the discussion in Section 5 to better reflect the breadth and recent advancements in integrating imaging data with other biomedical modalities. Specifically, we have added recent examples and citations that illustrate how integrating imaging with omic data can help discover biomarkers, improve prognostic accuracy, and enable earlier disease detection. We also highlighted the complementary roles of electronic health records (EHRs) and wearable devices in contextualising imaging data, contributing to a more comprehensive and time-resolved understanding of patient health. Discussion of recent advancements in techniques is dedicated to Section 5.1 and elaborated to

include MLLM and foundation models, as elaborated in the previous comment. Section 5.3 is dedicated to outlining future directions, explicitly discussing the integration of MLLMs with text and highlighting the advantages of text integration, such as improving contextual understanding, enhancing explainability and adding zero-shot learning capability.

The revised text from Section 5.3 is reported below.

“...For instance, integrating imaging with genomic data can reveal links between structural abnormalities and molecular profiles, which enhances understanding of disease mechanisms [137–139]. 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 [140–142]. 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 [143]. Beyond 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 [144]. 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 [145]. Additionally, text integration supports zero-shot learning by enabling the development of generalist models [146]. Integration of text with single imaging modalities, such as histopathology [145, 147, 148] or radiology [149–151], 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, 133, 152] to enable more comprehensive analyses and deeper insights into complex biological systems....”

- Finally, while the review offers a broad and informative overview, it would be helpful for the authors to articulate more clearly what distinguishes this work from existing review articles on multimodal biomedical imaging. Clarifying the novel insights or unique perspectives provided in this paper could better highlight its contribution to the field.

We thank the reviewer for pointing this out, which helped us clarify the contribution of our review. While an increasing number of reviews address multimodal applications, there remains a lack of comprehensive analyses focused on the latest AI-driven integration strategies for multimodal biomedical imaging. Our review specifically addresses this gap by providing an in-

depth examination of recent advances in AI methods tailored to image-to-image integration. The key points have been integrated into the main text (Section 1 on page 3) and are summarised below.

- **Specific focus on multimodal biomedical imaging, focusing on AI applications on image-to-image integration:** While many studies have reviewed the integration of various biomedical data types (including medical imaging, electronic health records (EHRs), and data from wearable or ambient biosensors [1]), they often treat imaging as just one component among others. Due to the heterogeneity of biomedical data in terms of format and structure, these reviews do not address the specific applications associated with image-to-image integration. Our review fills this gap by focusing specifically on AI-driven strategies for integrating image-to-image modalities.
- **Technical challenges specific to integrating imaging data:** Due to the unique scope of biomedical imaging, we address and discuss technical challenges unique to integrating imaging data (e.g., modality heterogeneity and spatial misalignment), often underexplored in general multimodal reviews. While most of the fusion strategies can be applied in many scenarios, some fusion strategies, such as pixel-level fusion, can only work in imaging integration, which needs precise spatial alignment. The existing reviews focus on the overall technical challenges when integrating multimodal data (e.g., wearable-derived data, omics data, and imaging data). However, they do not report any detailed discussion on the technical challenges and solutions when integrating imaging data generated from different sources.
- **Ethical considerations and interpretability of multimodal AI:** Our review incorporates ethical considerations alongside explainability in multimodal AI approaches for biomedical imaging, offering a complementary perspective to existing multimodal reviews. While several reviews discuss privacy-related aspects, they tend to focus on broader multimodal data and may not address the specific ethical and interpretability challenges associated with image-to-image integration. In this review, we go beyond privacy to discuss issues such as interpretability, bias, and accountability in the context of biomedical imaging, highlighting emerging solutions and outlining future research directions.

To reflect these contributions, we have highlighted the distinct perspectives offered by our review in comparison to existing literature. These points are now clearly presented in Section 1 (page 3) of the revised manuscript, as follows:

“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 data. Hence, this review specifically 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. Finally, to address these challenges, we present emerging solutions and outline potential future directions 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...”

References (used as part of our responses to Reviewer 2)

[1] Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, npj Digital Medicine, 2020.

Reviewer #3

This paper presents a thorough review of the biomedical image analysis pipeline when dealing with multimodal data. It covers the different stages from pre-processing, data augmentation, fusion and, finally, the downstream tasks. The paper is well-written and easy to follow.

Main comments

- The title is misleading. It gives the impression the paper deals with foundation models, which is not the case. Please revise

We thank the reviewer for this suggestion. We have revised the title of the manuscript by removing the direct reference to foundation models and focusing on multimodal integration in the biomedical imaging context. The updated title is ***“Bridging Modalities with AI: A Comprehensive Review of AI Innovations in Multimodal Biomedical Imaging”***.

- Multi-modal applications in biomedical imaging start to explore the usage of modalities beyond just images. New applications consider using text (reports) or electronic health records. This is a current trend [1] that should be presented in this review. Currently, it is just mentioned as a perspective

Thank you for raising this important point. We fully agree that multimodal biomedical modality, including imaging and other data types such as text and EHR, is an active area. We have expanded Section 5.3 (from page 25, line 743 to page 26, line 770) to discuss more of the use cases of integration with other data modalities. Specifically, Section 5.3 is devoted to future directions, explicitly discussing the integration of Multimodal Large Language Models (MLLMs) with text [2,3]. We highlight the advantages of text integration, such as improving contextual understanding, enhancing explainability and adding zero-shot learning capability. We also

discuss major challenges, including semantic alignment between modalities, particularly text tokens and image regions, and data heterogeneity.

The revised text from Section 5.3 is reported below.

“...For instance, integrating imaging with genomic data can reveal links between structural abnormalities and molecular profiles, which enhances understanding of disease mechanisms [137–139]. 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 [140–142]. 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 [143]. Beyond 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 [144]. 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 [145]. Additionally, text integration supports zero-shot learning by enabling the development of generalist models [146]. Integration of text with single imaging modalities, such as histopathology [145, 147, 148] or radiology [149–151], 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, 133, 152] to enable more comprehensive analyses and deeper insights into complex biological systems...”

- The paper's core, which should be about the joint usage of multiple image modalities, is relatively little discussed. Most of the paper is devoted to discussing image modalities, which is great but perhaps better suited for a textbook, and pre-processing, which is common to many other applications and not specific to multi-modal image analysis. It would be good to see a more in-depth study of data fusion, the existing applications, and their technical limitations.

We thank the reviewer for this remark. We have now revised the manuscript to add a more in-depth discussion on the strengths of multiple biomedical integrations in Section 1 (pages 3 and 4). Specifically, we have reduced the single modality discussion and placed more emphasis on the commentary information of biomedical imaging integration, as reported below.

“...Each biomedical imaging modality has unique characteristics, and 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 (Figure 1A), including X-rays, ultrasound, Computed Tomography (CT), and Magnetic Resonance Imaging (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 [13]. 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 suitable 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) [14]. Additionally, PET and 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, which complements the high-resolution anatomical detail offered by MRI and CT (e.g., localisation of lesions and identification of structural abnormalities) [15]. Hence, the integration of these images can combine the strengths of each imaging type to have a holistic picture of both structural and functional information, improving the precision of diagnoses [16]....

Additionally, we have also expanded the discussion on data fusion throughout Section 3 from page 15 to page 18, including the existing applications and their technical limitations. Specifically, we have included more existing applications of multimodal biomedical imaging and examined their technical challenges. This includes a comparative discussion on the performance of different fusion strategies under real-world conditions (like data misalignment and noise) to provide more practical perspectives on multiple fusion strategies.

Pixel-level fusion: “... Pixel-level fusion (also referred to as early fusion, Figure 2A(i)) combines different biomedical imaging modalities of the same sample region at a pixel-level to create a single and enhanced image [67]....”

“...For instance, combining pixel-level information from CT, MRI, and PET has been shown to enhance spatial and anatomical features of biomedical images [69-72]. 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 [73]. 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 [70, 74]....

...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...

...However, pixel-level fusion combines raw images and it can only be applied on 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, significantly 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 (See Section 2.1). 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 [78]...

Feature-level fusion: “Feature-level fusion (also referred to as intermediate or joint fusion, Figure 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 [66]. 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 [66]. 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...

...Additionally, feature-level fusion also offers more flexibility in the regions of interest to be fused. 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 or structures within the body, enabling a more holistic understanding of the patient’s condition... Histopathological imaging has been used to integrate 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 provides a more comprehensive perspective and enhances the clinical diagnosis [21]...”

“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, Figure 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 [66]. Moreover, by combining high-level decisions inferred from individual modalities, this fusion strategy has shown robust results in many applications and can handle the data misalignment and noise challenges between multiple imaging modalities. Decision-level fusion can integrate predictions from different biomedical imaging

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 subtype identification. ...Moreover, decision-level fusion has also significantly improved cancer patients' outcomes by integrating biomedical imaging from different biological views, such as radiomics from CT scans and histopathological features from WSI [25]....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, 89]...”.

“...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 to train 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 (Figure 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 [66]...”

“...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... Despite its promising performance, combining multiple abstraction levels requires complex architectures, resulting in significantly 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 [93]), achieving effective optimisation and information fusion across heterogeneous modalities in hierarchical-level fusion remains a significant 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 significant improvements in model performance and flexibility, it still requires careful consideration of the trade-offs between model performance and complexity [66]...”

References (used as part of the responses to Reviewer 3)

- [1] Pérez-García, Fernando, et al. "Exploring scalable medical image encoders beyond text supervision." *Nature Machine Intelligence* (2025): 1-12
- [2] Lu, Ming Y., et al. "A multimodal generative AI copilot for human pathology." *Nature* 634.8033 (2024): 466-473.
- [3] Tu, Tao, et al. "Towards generalist biomedical ai, 2023." URL <https://arxiv.org/abs/2307.14334> (2023).

References (used when reporting the text copied from the manuscript)

- [5] Park, J., Choi, S., Knieling, F., Clingman, B., Bohndiek, S., Wang, L.V., Kim, C.: Clinical translation of photoacoustic imaging. *Nature Reviews Bioengineering*, 1–20 (2024)
- [6] Zhao, C., Guo, L., Dong, J., Cai, Z.: Mass spectrometry imaging-based multi-modal technique: Next-generation of biochemical analysis strategy. *The Innovation* 2(4) (2021)
- [7] Duan, J., Xiong, J., Li, Y., Ding, W.: Deep learning based multimodal biomedical data fusion: An overview and comparative review. *Information Fusion*, 102536 (2024)
- [8] Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. *Nature Communications* 15(1), 654 (2024)
- [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. *Journal of Medical and Biological Engineering*, 1–13 (2024)
- [11] Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. *Nature medicine* 29(8), 1930–1940 (2023)
- [14] Chen, J., Chen, J.: Multimodal image feature fusion for improving medical ultra- sound image segmentation. *Biomedical Signal Processing and Control* 89, 105705 (2024).
- [15] Ottestad, A.L., Johansen, H., Halvorsen, T.O., Dai, H.Y., Wahl, S.G.F., Emdal, E.F., Grønberg, B.H.: 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(1), 646 (2023)
- [16] 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(13), 3267 (2023)
- [19] Doherty, T., McKeever, S., Al-Attar, N., Murphy, T., Aura, C., Rahman, A., O'Neill, A., Finn, S.P., Kay, E., Gallagher, W.M., et al.: Feature fusion of raman chemical imaging and digital histopathology using machine learning for prostate cancer detection. *Analyst* 146(13), 4195–4211 (2021)
- [20] Nunes, J.B., Ijsselsteijn, M.E., Abdelaal, T., Ursem, R., Ploeg, M., Giera, M., Everts, B., Mahfouz, A., Heijs, B., Miranda, N.F.: Integration of mass cytometry and mass spectrometry imaging for spatially resolved single-cell metabolic profiling. *Nature Methods* 21(10), 1796–1800 (2024)
- [21] Wang, D., Liu, C., Wang, X., Liu, X., Lan, C., Zhao, P., Cho, W.C., Graeber, M.B., Liu, Y.: Automated machine-learning framework integrating histopathological and radiological information for predicting idh1 mutation status in glioma. *Frontiers in Bioinformatics* 1, 718697 (2021)
- [22] Fitzgerald, S., Akhtar, J., Schartner, E., Ebendorff-Heidepriem, H., Mahadevan- Jansen,

- A., Li, J.: Multimodal raman spectroscopy and optical coherence tomography for biomedical analysis. *Journal of biophotonics* 16(3), 202200231 (2023)
- [23] Liu, Y., Valji, K., Monsky, W., Zheng, C., Yang, X.: Optical imaging guidance in oncologic surgery and interventional oncology. *Pharmacological Research*, 107612 (2025)
- [24] Lin, L., Wang, L.V.: The emerging role of photoacoustic imaging in clinical oncology. *Nature Reviews Clinical Oncology* 19(6), 365–384 (2022)
- [25] Boehm, K.M., Aherne, E.A., Ellenson, L., Nikolovski, I., Alghamdi, M., V´azquez-García, I., Zamarin, D., Long Roche, K., Liu, Y., Patel, D., et al.: Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nature cancer* 3(6), 723–733 (2022)
- [26] Zhang, X., Zhang, Y., Zhang, G., Qiu, X., Tan, W., Yin, X., Liao, L.: Deep learning with radiomics for disease diagnosis and treatment: challenges and potential. *Frontiers in oncology* 12, 773840 (2022)
- [28] Cammarasana, S., Nicolardi, P., Patan`e, G.: Real-time denoising of ultrasound images based on deep learning. *Medical & Biological Engineering & Computing* 60(8), 2229–2244 (2022)
- [29] Chen, H., Zhang, Y., Kalra, M.K., Lin, F., Chen, Y., Liao, P., Zhou, J., Wang, G.: Low-dose ct with a residual encoder-decoder convolutional neural network. *IEEE transactions on medical imaging* 36(12), 2524–2535 (2017)
- [30] Gao, Q., Shan, H.: Cocosdiff: a contextual conditional diffusion model for low- dose ct image denoising. In: *Developments in X-Ray Tomography XIV*, vol. 12242, pp. 92–98 (2022). SPIE
- [31] Lin, G., Jin, Y., Huang, Z., Chen, Z., Liu, H., Zhou, C., Zhang, X., Fan, W., Zhang, N., Liang, D., et al.: Multimodal feature-guided diffusion model for low- count pet image denoising. *Medical Physics* (2025)
- [32] Parvaiz, A., Khalid, M.A., Zafar, R., Ameer, H., Ali, M., Fraz, M.M.: Vision transformers in medical computer vision—a contemplative retrospection. *Engineering Applications of Artificial Intelligence* 122, 106126 (2023)
- [33] Singh, A.R., Athisayamani, S., Karim, F.K., Ibrahim, A.Z., Alshetewi, S., Mostafa, S.M.: An enhanced denoising system for mammogram images using deep transformer model with fusion of local and global features. *Scientific Reports* 15(1), 6562 (2025)
- [34] Song, T.-A., Yang, F., Dutta, J.: Noise2void: unsupervised denoising of pet images. *Physics in Medicine & Biology* 66(21), 214002 (2021)
- [35] Shen, Y., Chen, L., Liu, J., Chen, H., Wang, C., Ding, H., Zhang, Q.: Pads-net: Gan-based radiomics using multi-task network of denoising and segmentation for ultrasonic diagnosis of parkinson disease. *Computerized Medical Imaging and Graphics* 120, 102490 (2025)
- [36] Kim, S., Al-masni, M.A., Lee, S., Jung, S., Jung, K.-J., Cui, C., Gho, S.- M., Choi, Y.H., Kim, D.-H.: Unsupervised learning for motion correction and assessment in brain magnetic resonance imaging using severity-based regularized cycle consistency. *Engineering Applications of Artificial Intelligence* 142, 109978 (2025)
- [37] Guo, K., Chen, J., Qiu, T., Guo, S., Luo, T., Chen, T., Ren, S.: Medgan: An adaptive gan approach for medical image generation. *Computers in Biology and Medicine* 163, 107119 (2023)

- [38] 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. *Biomedical Signal Processing and Control* 81, 104484 (2023)
- [39] Chen, Z., Yoon, S., Strotzer, Q., Khalid, R.N., Tivnan, M., Li, Q., Gupta, R., Wu, D.: Portable head ct motion artifact correction via diffusion-based generative model. *Computerized Medical Imaging and Graphics* 119, 102478 (2025)
- [41] Xu, Y., Quan, R., Xu, W., Huang, Y., Chen, X., Liu, F.: Advances in medical image segmentation: a comprehensive review of traditional, deep learning and hybrid approaches. *Bioengineering* 11(10), 1034 (2024)
- [42] Zhou, T., Ruan, S., Canu, S.: A review: Deep learning for medical image segmentation using multi-modality fusion. *Array* 3, 100004 (2019)
- [43] Abualigah, L., Almotairi, K.H., Elaziz, M.A.: Multilevel thresholding image segmentation using meta-heuristic optimization algorithms: Comparative analysis, open challenges and new trends. *Applied Intelligence* 53(10), 11654–11704 (2023)
- [44] Fernandez, V., Pinaya, W.H.L., Borges, P., Tudosiu, P.-D., Graham, M.S., Vercauteren, T., Cardoso, M.J.: Can segmentation models be trained with fully synthetically generated data? In: *International Workshop on Simulation and Synthesis in Medical Imaging*, pp. 79–90 (2022). Springer
- [45] Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* 18(2), 203–211 (2021)
- [46] Zhuang, X., Xu, J., Luo, X., Chen, C., Ouyang, C., Rueckert, D., Campello, V.M., Lekadir, K., Vesal, S., RaviKumar, N., et al.: Cardiac segmentation on late gadolinium enhancement mri: a benchmark study from multi-sequence cardiac mr segmentation challenge. *Medical Image Analysis* 81, 102528 (2022)
- [47] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: *International MICCAI Brain lesion Workshop*, pp. 272–284 (2021). Springer
- [48] Xiao, H., Li, L., Liu, Q., Zhu, X., Zhang, Q.: Transformers in medical image segmentation: A review. *Biomedical Signal Processing and Control* 84, 104791 (2023)
- [49] 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 Analytics*, 100340 (2024)
- [50] Dayarathna, S., Islam, K.T., Uribe, S., Yang, G., Hayat, M., Chen, Z.: Deep learning based synthesis of mri, ct and pet: Review and analysis. *Medical image analysis* 92, 103046 (2024)
- [51] Nag, A., Mondal, H., Hassan, M.M., Al-Shehari, T., Kadrie, M., Al-Razgan, M., Alfakih, T., Biswas, S., Bairagi, A.K.: Tumorganet: A transfer learning and generative adversarial network-based data augmentation model for brain tumor classification. *IEEE Access* (2024)
- [52] Guan, Q., Chen, Y., Wei, Z., Heidari, A.A., Hu, H., Yang, X.-H., Zheng, J., Zhou, Q., Chen, H., Chen, F.: Medical image augmentation for lesion detection using a texture-constrained multichannel progressive gan. *Computers in Biology and Medicine* 145, 105444 (2022)
- [53] Pombo, G., Gray, R., Cardoso, M.J., Ourselin, S., Rees, G., Ashburner, J., Nachev, P.:

Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3d deep generative models. *Medical Image Analysis* 84, 102723 (2023)

[54] Zhang, Y., Peng, C., Wang, Q., Song, D., Li, K., Zhou, S.K.: Unified multi-modal image synthesis for missing modality imputation. *IEEE Transactions on Medical Imaging* (2024)

[55] Saad, M.M., O'Reilly, R., Rehmani, M.H.: A survey on training challenges in generative adversarial networks for biomedical image analysis. *Artificial Intelligence Review* 57(2), 19 (2024)

[56] Han, C., Rundo, L., Araki, R., Nagano, Y., Furukawa, Y., Mauri, G., Nakayama, H., Hayashi, H.: Combining noise-to-image and image-to-image gans: Brain mr image augmentation for tumor detection. *Ieee Access* 7, 156966–156977 (2019)

[57] 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 bioinformatics* 26(1), 29 (2025)

[58] Qiang, N., Dong, Q., Liang, H., Ge, B., Zhang, S., Sun, Y., Zhang, C., Zhang, W., Gao, J., Liu, T.: Modeling and augmenting of fmri data using deep recurrent variational auto-encoder. *Journal of neural engineering* 18(4), 0460–6 (2021)

[59] Huo, J., Vakharia, V., Wu, C., Sharan, A., Ko, A., Ourselin, S., Sparks, R.: Brain lesion synthesis via progressive adversarial variational auto-encoder. In: *International Workshop on Simulation and Synthesis in Medical Imaging*, pp. 101–111 (2022). Springer

[60] Wei, R., Mahmood, A.: Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey. *Ieee Access* 9, 4939–4956 (2020)

[61] Neloy, A.A., Turgeon, M.: A comprehensive study of auto-encoders for anomaly detection: Efficiency and trade-offs. *Machine Learning with Applications*, 100572 (2024)

[62] Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacıhaliloglu, I., Merhof, D.: Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis* 88, 102846 (2023)

[63] Wang, Y., Hu, H., Yu, S., Yang, Y., Guo, Y., Song, X., Chen, F., Liu, Q.: A unified hybrid transformer for joint mri sequences super-resolution and missing data imputation. *Physics in Medicine & Biology* 68(13), 135006 (2023)

[64] Liu, J., Pasumarthi, S., Duffy, B., Gong, E., Datta, K., Zaharchuk, G.: One model to synthesize them all: Multi-contrast multi-scale transformer for missing data imputation. *IEEE transactions on medical imaging* 42(9), 2577–2591 (2023)

[65] Azad, R., Dehghanmanshadi, M., Khosravi, N., Cohen-Adad, J., Merhof, D.: Addressing missing modality challenges in mri images: A comprehensive review. *Computational Visual Media* 11(2), 241–268 (2025)

[66] Kalamkar, S., et al.: Multimodal image fusion: A systematic review. *Decision Analytics Journal*, 100327 (2023)

[69] Liu, J., Dian, R., Li, S., Liu, H.: Sgffusion: A saliency guided deep-learning framework for pixel-level image fusion. *Information Fusion* 91, 205–214 (2023)

[70] Zhou, J., Xing, X., Yan, M., Yuan, D., Zhu, C., Zhang, C., Xu, T.: A fusion algorithm based on composite decomposition for pet and mri medical images. *Biomedical Signal Processing and Control* 76, 103717 (2022)

[71] Li, X., Zhou, F., Tan, H., Zhang, W., Zhao, C.: Multimodal medical image fusion based on joint bilateral filter and local gradient energy. *Information Sciences* 569, 302–325 (2021)

- [72] Odusami, M., Maskeliūnas, R., Damaševičius, R.: Pixel-level fusion approach with vision transformer for early detection of alzheimer's disease. *Electronics* 12(5), 1218 (2023)
- [73] Bischof, J., Fletcher, G., Verkade, P., Kuntner, C., Fernandez-Rodriguez, J., Chaabane, L., Rose, L.A., Walter, A., Vandenbosch, M., Zandvoort, M.A., et al.: Multimodal bioimaging across disciplines and scales: challenges, opportunities and breaking down barriers. *npj Imaging* 2(1), 5 (2024)
- [74] Yuan, Y., Cheng, Y.: Medical image segmentation with unet-based multi-scale context fusion. *Scientific Reports* 14(1), 15687 (2024)
- [78] Bierbrier, J., Gueziri, H.-E., Collins, D.L.: Estimating medical image registration error and confidence: A taxonomy and scoping review. *Medical Image Analysis* 81, 102531 (2022)
- [89] Zhang, S., Metaxas, D.: On the challenges and perspectives of foundation models for medical image analysis. *Medical image analysis* 91, 102996 (2024)
- [93] Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: Development of a large-scale medical visual question-answering dataset. *Communications Medicine* 4(1), 277 (2024)
- [97] 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, pp. 317–325 (2020). Springer
- [129] Pai, S., Bontempi, D., Hadzic, I., Prudente, V., Sokač, M., Chaunzwa, T.L., Bernatz, S., Hosny, A., Mak, R.H., Birkbak, N.J., et al.: Foundation model for cancer imaging biomarkers. *Nature machine intelligence* 6(3), 354–367 (2024)
- [130] Perez-Lopez, R., Ghaffari Laleh, N., Mahmood, F., Kather, J.N.: A guide to artificial intelligence for cancer researchers. *Nature Reviews Cancer* 24(6), 427–441 (2024)
- [131] Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., Merhof, D.: Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689* (2023)
- [132] Wang, L., Wang, H., Yang, H., Mao, J., Yang, Z., Shen, J., Li, X.: Interpretable bilingual multimodal large language model for diverse biomedical tasks. *arXiv preprint arXiv:2410.18387* (2024)
- [133] Huang, X., Shen, L., Liu, J., Shang, F., Li, H., Huang, H., Yang, Y.: Towards a multimodal large language model with pixel-level insight for biomedicine. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 3779–3787 (2025)
- [137] Qoku, A., Katsaouni, N., Flinner, N., Buettner, F., Schulz, M.H.: Multi-modal analysis methods in predictive biomedicine. *Computational and structural biotechnology journal* (2023)
- [138] Wu, J., Chen, Y., Wang, P., Caselli, R.J., Thompson, P.M., Wang, J., Wang, Y.: Integrating transcriptomics, genomics, and imaging in alzheimer's disease: A federated model. *Frontiers in radiology* 1, 777030 (2022)
- [139] Adewale, Q., Khan, A.F., Carbonell, F., Iturria-Medina, Y., Initiative, A.D.N.: Integrated transcriptomic and neuroimaging brain model decodes biological mechanisms in aging and alzheimer's disease. *Elife* 10, 62589 (2021)

- [140] Pirruccello, J.P., Chaffin, M.D., Chou, E.L., Fleming, S.J., Lin, H., Nekoui, M., Khurshid, S., Friedman, S.F., Bick, A.G., Arduini, A., et al.: Deep learning enables genetic analysis of the human thoracic aorta. *Nature genetics* 54(1), 40–51 (2022)
- [141] Ning, C., Fan, L., Jin, M., Wang, W., Hu, Z., Cai, Y., Chen, L., Lu, Z., Zhang, M., Chen, C., et al.: Genome-wide association analysis of left ventricular imaging-derived phenotypes identifies 72 risk loci and yields genetic insights into hypertrophic cardiomyopathy. *Nature Communications* 14(1), 7900 (2023)
- [142] Gomes, B., Singh, A., O’Sullivan, J.W., Schnurr, T.M., Goddard, P.C., Loong, S., Amar, D., Hughes, J.W., Kostur, M., Haddad, F., et al.: Genetic architecture of cardiac dynamic flow volumes. *Nature Genetics* 56(2), 245–257 (2024)
- [143] Yu, Y., Ren, W., Mao, L., Ouyang, W., Hu, Q., Yao, Q., Tan, Y., He, Z., Ban, X., Hu, H., et al.: Mri-based multimodal ai model enables prediction of recurrence risk and adjuvant therapy in breast cancer. *Pharmacological Research*, 107765 (2025)
- [144] Zhang, S., Xiao, X., Yi, Y., Wang, X., Zhu, L., Shen, Y., Lin, D., Wu, C.: Tumor initiation and early tumorigenesis: molecular mechanisms and interventional targets. *Signal transduction and targeted therapy* 9(1), 149 (2024)
- [145] Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Zhao, M., Chow, A.K., Ikemura, K., Kim, A., Pouli, D., Patel, A., et al.: A multimodal generative ai copilot for human pathology. *Nature* 634(8033), 466–473 (2024)
- [146] Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. *Nature* 616(7956), 259–265 (2023)
- [147] Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. *Nature Medicine* 30(3), 863–874 (2024)
- [148] Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. *Nature Medicine* 30(3), 850–862 (2024)
- [149] Rao, V.M., Hla, M., Moor, M., Adithan, S., Kwak, S., Topol, E.J., Rajpurkar, P.: Multimodal generative ai for medical image interpretation. *Nature* 639(8056), 888–896 (2025)
- [150] P´erez-Garc´ıa, F., Sharma, H., Bond-Taylor, S., Bouzid, K., Salvatelli, V., Ilse, M., Bannur, S., Castro, D.C., Schwaighofer, A., Lungren, M.P., et al.: Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 1–12 (2025)
- [151] Lee, J.-O., Zhou, H.-Y., Berzin, T.M., Sodickson, D.K., Rajpurkar, P.: Multi-modal generative ai for interpreting 3d medical images and videos. *npj Digital Medicine* 8(1), 1–8 (2025)
- [152] Tu, T., Azizi, S., Driess, D., Schaeckermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al.: Towards generalist biomedical ai, 2023. URL <https://arxiv.org/abs/2307.14334> (2023)

Reviewer #2 (Remarks to the Author):

Thanks to the author for providing an extensive response to address my questions and concerns. I especially appreciate the author's efforts to revise Section 3 and 5, which have been improved a lot. Below is a follow-up questions with remaining concerns after reading the revised manuscript and point-to-point response.

We thank the reviewer for their feedback. Below, we provide the detailed responses to their questions.

1. The new categories look more reasonable, while it may still need more thinking about what is “data-driven imaging”. It is not the image data collected from a physical scanner, while processed by some techniques to extract information. Is it a specific imaging modality or should it belong to an image processing procedure? It may be confusing for the reader since there is another section later to discuss about the image processing techniques.

We appreciate the reviewer's suggestions about the “data-driven imaging”. We agree that “data-driven imaging”, which refers to information extracted from the biomedical image, is more appropriately categorised under image processing or analytical tasks rather than as a standalone imaging modality.

To address this point, we have now (i) removed “data-driven imaging” from Box 1: “**Bioimaging Modalities**” and (ii) moved the related discussion from Section 1.1 to Section 3.2 “**Downstream tasks and biological insights**” on page 18, where it more appropriately fits within the context of image analysis and interpretation. We copied below the relevant section.

“...Next, data-driven imaging, including radiomics and radiogenomics, represents a significant advancement in the biomedical field, driven by AI innovations to provide more insights into the disease. Radiomics derives tumour information (e.g., tumour texture, shape, and intensity) from radiological images, and radiogenomics aims to correlate 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 disease. With the support of AI, radiomics and radiogenomics are 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 [78, 108, 109]...”

The imaging modalities listed and discussed in the updated Introduction and Box 1 focus now on three main categories only (i.e., radiological imaging, microscopic and spectroscopic imaging, and optical imaging). The revised Box 1 is reported on page 5 and also copied below.

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.

2. After the revision, the structure and rationale of Section 2 still look kind of weird. It is not clear what is the highlight or key point in this section: to introduce different image processing tasks? Or different models? Or focus on multimodal images? It may be confusing to claim in this section: “we explore how several AI-driven techniques are applied in preprocessing the multi-modal biomedical images (Figure 1B and C) to improve image quality, handle missing modalities and identify regions of interest.” Many of the works discussed in this section are only limited to only preprocessing but also downstream tasks. And how the missing modalities connected to other tasks are not clear either. This may also lead to confusion to understand Table 1. What is the key message or conclusion the paper expects the reader to extract from reading this long list of papers? It may be helpful to think further about it.

We thank the reviewer for this valuable and constructive feedback. We agree that distinguishing imaging processing from downstream tasks would further clarify our text. We also agree that some of the processing tasks, such as segmentation, may not be limited to the preprocessing scope.

Following the reviewer's suggestions, we have significantly revised Section 2 to improve its coherence and purpose. Specifically:

- We clarified the scope and focus of Section 2, which now focuses on AI-driven image processing tasks only, including Subsection 2.1 - ***“Biomedical image denoising and artefact correction”*** and Subsection 2.2 - ***“Generative strategies for data augmentation”***.
- We removed the “missing modalities” discussions from this Section and focused on data augmentation to maintain consistency with other imaging processing tasks.
- We discuss AI-driven techniques in each task and the potential of these tasks in improving multimodal analysis.
- We also removed segmentation-related content from Section 2 and relocated it to Section 3, where we discuss segmentation in the context of downstream applications. Section 3.2 has now been updated as follows:

“...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. 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 [87, 88]. These models are particularly effective for MRI due to their ability to model complex anatomical structures....

... 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 [102], 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 significant step towards universal biomedical segmentation solutions [8]...

- We also moved the segmentation-related entries from Table 1 and added them to Table 2, since they were also related to different fusion techniques with the associated downstream task.

- Additionally, we have renamed Section 2 and Table 1 to reflect their focus and scope as Section 2: “**AI-driven biomedical image processing methods**” and Table 1: “**AI-driven processing techniques for biomedical imaging**”.

Overall, in the revised manuscript, Section 2 now highlights the main AI-driven processing techniques in biomedical imaging, supported by Table 1 that summarises the main papers and techniques in the field, while Section 3 focuses on multimodal integration strategies and downstream tasks.

3. The author claims some key contributions that make this review distinctive from previous review papers. However, in this active field, there have been quite a few review articles which may have covered the several points claimed by the author to some extent. To just name a few as the list below. Thus, I am afraid that it may not be very convincing to claim these as unique contributions in the paper.

[1] Vision-language foundation models for medical imaging: a review of current practices and innovations, Biomedical Engineering Letters, 2025

[2] Multimodal Machine Learning in Image-Based and Clinical Biomedicine: Survey and Prospects, IJCV, 2024

[3] Advancements in Medical Radiology Through Multimodal Machine Learning: A Comprehensive Overview, Bioengineering, 2025

[4] A systematic review of intermediate fusion in multimodal deep learning for biomedical applications, Image and Vision Computing, 2025

[5] Foundation Model for Advancing Healthcare: Challenges, Opportunities and Future Directions, IEEE Rev Biomed Eng., 2025

We thank the reviewer for bringing up this important point and for highlighting several recently published, relevant, and high-quality review articles in the field. We agree that multimodal biomedical imaging is a rapidly evolving area, and we acknowledge that some of the topics we discuss, such as fusion strategies and the role of foundation models, have indeed been covered to some extent in recent surveys.

We have included these references in our updated manuscript and clarified in Section 1, as detailed below, the distinct contributions of our review compared with similar works.

While [1] and [5] focus primarily on foundation and vision-language foundation models, our review provides a broader overview of imaging fusion strategies (e.g., pixel, feature, decision, and hierarchical-level approaches) for integrating multiple types of biomedical imaging (e.g., radiological, microscopic, and spectroscopic modalities) and highlights their advantages and limitations. In contrast, [2] offers a wider survey of multimodal machine learning across both imaging- and non-imaging-based data sources (e.g., EHRs and genomics), whereas our work emphasises integration strategies and technical issues specifically tailored to biomedical imaging.

We also provide a detailed discussion of how complementary structural, functional, and molecular imaging can be exploited to improve clinical decision-making, and we outline emerging directions such as foundation models and AI agents.

On the other hand, while [3] concentrates on advances in machine learning applied to radiological imaging, we expand the scope to include histopathology, microscopy, spectroscopy, and imaging–omics fusion. Similarly, although [4] systematically analyses intermediate fusion methods, we highlight several integration strategies for biomedical applications and emphasise how complementary imaging information can be harnessed to provide more insights into underlying biological mechanisms and improve diagnostic and prognostic tasks. Taken together, our review complements these existing works by offering a holistic and imaging-specific review that bridges technical advances, clinical applications, and future opportunities in multimodal biomedical imaging.

Overall, four main unique contributions of the proposed review can be highlighted:

1. **Focus on imaging-specific integration challenges:** We address several technical challenges that are unique to integrating biomedical imaging data (e.g., modality heterogeneity, image enhancement, spatial misalignment, and data augmentation methods), which are often underexplored or only briefly mentioned in general multimodal AI reviews.
2. **Modality-aware integration strategies:** Beyond technically examining and discussing several integration strategies to merge multiple biomedical imaging types, we highlighted that the complementary information across several imaging modalities (e.g., structural, functional, and molecular data) can be harnessed to improve clinical decision-making, offering deeper insights into the synergistic potential of multimodal imaging. Also, we identified key challenges and proposed potential solutions to improve data fusion (based on biomedical imaging type and resolutions), model performance, and clinical applicability.
3. **Ethical and interpretability aspects:** We offered a unique perspective on interpretability and ethical challenges and presented emerging solutions.
4. **Perspective:** Rather than a systematic review, we outlined potential future directions, such as MLLMs and AI agents, to further enhance the successful applications of AI in multimodal biomedical imaging and guide continuous innovation in this field. We also added the timeline infographic of key AI developments in the field in the past fifteen years, highlighting the innovations and growing role of multimodal integration in biomedical imaging

Following the reviewer's suggestion, the contribution of this work has been better highlighted in the revised manuscript in Section 1 on pages 2 and 3, while also considering the recently published reviews.

The revised text added in Section 1 is reported below.

“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 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., 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.”

Reviewer #3 (Remarks to the Author):

The authors have undertaken a substantial amount of work to improve the manuscript. The new version is more solid and complete.

We thank the reviewer for their constructive comments. The point-by-point responses are reported below.

I recommend, however, revising Section 3.2 regarding the downstream tasks, as it remains weak. As this is a review paper, it should be motivated by an important amount of work in a domain (multi-modal biomedical imaging) that requires a survey summarizing the main achievements of the last couple of years (sometimes decades).

When reading Section 3.2, one gets the impression that the field has not reached a critical mass to "deserve" such a survey. While some works are cited, these mainly cover multi-modal imaging applications that have been long established, such as multi-sequence MR in cardiac and brain imaging, or the combination of US and MR in cancer imaging.

We appreciate the reviewer's suggestion regarding Section 3.2 and agree that this section could benefit from a broader and more forward-looking perspective on downstream tasks in multimodal biomedical imaging.

To address this, we have significantly expanded Section 3.2 (pages 17 – 23), adding the points below:

1. **Broaden the scope of the downstream task.** We have expanded the discussion on segmentation and included a wider range of downstream applications beyond conventional multimodal combinations (e.g., multi-sequence MRI). This includes examples of the emerging applications:
 - a. Microscopic and spectroscopic imaging (e.g., histopathology, Raman spectroscopy, hyperspectral imaging)
 - b. Optical imaging modalities (e.g., OI and OCT)
 - c. Other emerging applications, such as multimodal integration of imaging with clinical records, text, and omics data.
2. **Highlight the role of the emerging AI model in multimodal biomedical imaging integration.** We have also added a discussion on how foundation models (e.g., MedSAM, BiomedParse, BiomedGPT) are transforming downstream analysis in multimodal biomedical imaging by enabling better generalisation, scalability, and cross-modal understanding.

The revised text is reported below and in Section 3.2 (pages 17 - 23):

"... 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 [55, 92]. These models are particularly effective for MRI due to their ability to model complex anatomical structures...

... 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 [73, 95, 98]. 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 enable the generation of virtual histological stains to reduce the cost and variability of conventional chemical staining and conserve precious tissue samples [99, 100]...

... 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 [24, 101]. 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 [102]. Moreover, OCT-derived structural features, including the retinal nerve fibre layer and the ganglion cell–inner plexiform layer, have 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 [103].

Next, data-driven imaging, including radiomics and radiogenomics, represents a significant advancement in the biomedical field, driven by AI innovations to provide more insights into the disease. Radiomics derives tumour information (e.g., tumour texture, shape, and intensity) from radiological images, and radiogenomics aims to correlate 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 disease. With the support of AI, radiomics and radiogenomics are 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 [76, 104, 105].

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), electronic health records (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 [107]. For instance, integrating imaging with genomic data can reveal links between structural abnormalities and molecular profiles, which enhances understanding of disease

mechanisms [108–110]. 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 [111–113]. 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 [114]. Beyond 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 [115].

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 enabling a more complete and time-resolved understanding of patient health...”

“...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 [106], 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 significant step towards universal biomedical segmentation solutions [8].

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 [116]. Additionally, text integration supports zero-shot learning by enabling the development of generalist models [117]. Integration of text with single imaging modalities, such as histopathology [116, 118, 119] or radiology [120–122], 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 [123–125] to enable more comprehensive analyses and deeper insights into complex biological systems...”

“...In summary, the advances outlined above reflect the AI innovations in biomedical imaging over the past fifteen years, as illustrated in Figure 3. The field was initially dominated by traditional machine learning approaches 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 [126] in the ImageNet challenge has opened a new era of applying deep learning, particularly CNNs and pretrained models, to biomedical imaging [127]. Architectures, such as U-Net [128], GANs [129], and diffusion models [130], subsequently revolutionised the applications of AI, enabling models to focus on ROIs, 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 [131] and facilitated advanced analyses in cell and molecular signatures from high-resolution imaging (e.g., microscopic and spectroscopic) [132]. 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 [133] 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 [134], 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, 135, 136]. 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...”

To make their point, the authors should consider expanding the section while providing concrete examples of emerging applications that go beyond classical ones, where novel multi-modal biomedical images are considered. A table summarizing these, would also be of great value.

We thank the reviewer for this suggestion. In the revised manuscript, Section 3.2 has been considerably expanded as discussed in our response above.

In response to the reviewer's suggestion to summarise the main achievements of recent years, we created an infographic highlighting key advances. This is included as Figure 3 on page 21 of the revised manuscript and is also copied below.

The timeline infographic summarises key achievements and milestones over the past fifteen years in the multimodal AI and imaging field, helping contextualise the increasing relevance of downstream applications and providing context for the discussion presented in Section 3.2.

Specifically, the figure presents the evolution of AI in biomedical imaging from early machine learning approaches on single modalities (in 2010) to the emergence of multimodal foundation models and AI agents (in 2025 and forward). Initial progress was constrained by limited computational resources and datasets, but breakthroughs such as AlexNet in 2012 enabled deep learning methods (CNNs, pre-trained networks) to be adapted for biomedical tasks. Since 2016, multimodal integration has accelerated with vision–language models (e.g., CLIP) extending applications to personalised medicine and real-time monitoring. More recently, foundation models and multimodal large language models trained on imaging, text, and EHRs have demonstrated strong generalisation across tasks, paving the way for AI agents that integrate seamlessly with clinical ecosystems. Looking ahead, specialist multimodal healthcare agents are expected to unify imaging, wearables, and EHRs to support real-time clinical decision-making, provided these innovations remain transparent, ethical, and aligned with regulatory standards.

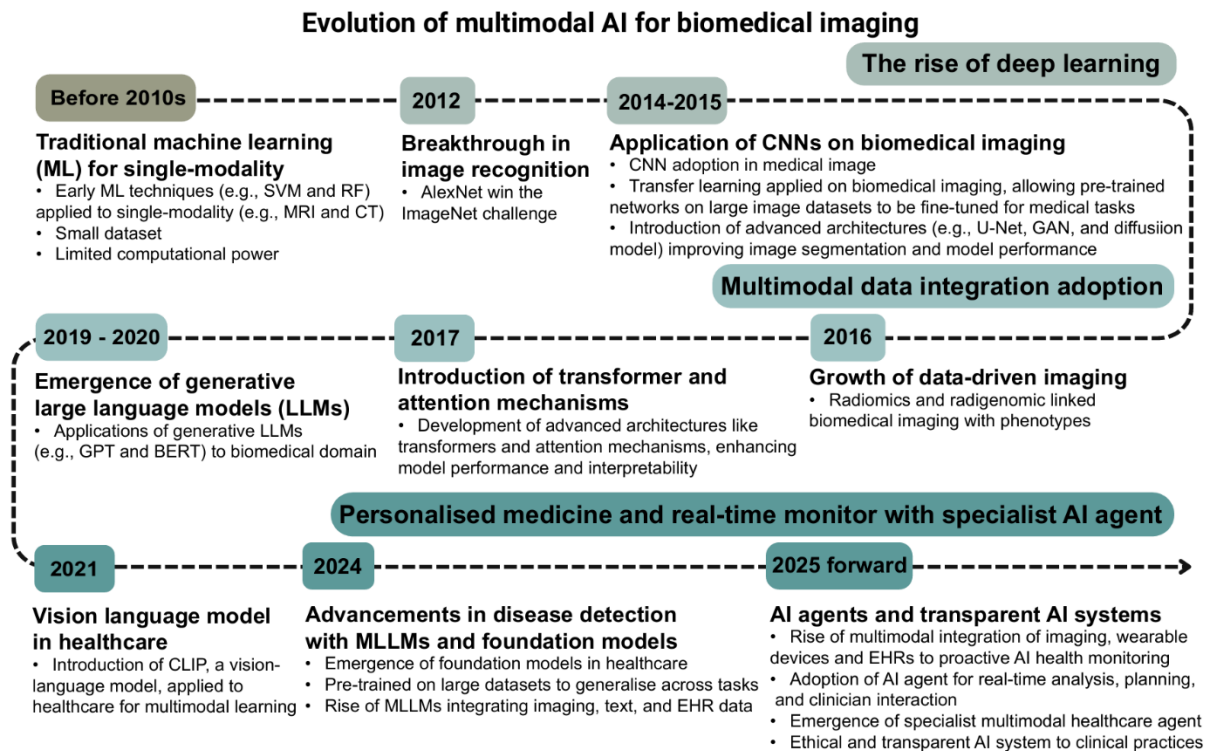


Figure 3: The evolution of AI in multimodal biomedical imaging over the past fifteen years. Early approaches relied on traditional machine learning (e.g., Support Vector Machine (SVM) and Random Forest (RF)) applied to a single imaging modality. The adoption of AI in the biomedical imaging field remains limited due to insufficient computational power and the scarcity of large, publicly available datasets. 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 such as GPT and BERT, and extended AI applications to the biomedical domain with natural language generation and mining. The success vision language model (e.g., CLIP) in healthcare has 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, text and EHR data, have demonstrated the capacity to generalise across several downstream tasks without the need for task-specific architectures. This has provided an opportunity for the emergence of AI-agents to integrate AI models with the clinical bioimaging ecosystem, including biomedical databases, diagnostic tools, and analytical software, to improve the real-time clinical decision-making process. Looking ahead, the rise of AI agents, especially specialist multimodal healthcare agents, is anticipated to streamline workflows by integrating imaging, wearable devices, and EHRs, and provide real-time analysis and interaction between clinicians and AI. Importantly, these AI innovations must be accompanied by a strong emphasis on ethical and transparent systems, ensuring that clinical practices benefit from systems that are explainable, trustworthy, and aligned with regulatory standards.

Point-by-point Response

Reviewer #2 (Remarks to the Author):

Thank the author for the exhausted and detailed responses to answer my questions and address the concerns. Most of my concerns have been well addressed while there are a few follow-up points that may need further attention from the author.

- Q1: The author is encouraged to reconsider more about the category of “Data-driven imaging”, since it is actually not a “physical imaging” modality but an image or data processing technique that can be essentially different from the other imaging categories. It may be more reasonable to be discussed in Section 2.

Thank you for your suggestion. We have moved the data-driven imaging category to Section 2 as suggested and updated the relevant content in Section 2.3 on page 10 of the marked manuscript.

The revised text added in **Section 2.3 - Feature extraction and data-driven imaging** is reported below:

“...In addition to traditional image processing techniques, data-driven imaging represents a significant advancement in the biomedical field, driven by AI innovations to provide more insights into the disease. 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.

Data-driven imaging includes radiomics and radiogenomics. Radiomics derives tumour information (e.g., tumour texture, shape, and intensity) from radiological images, and radiogenomics aims to correlate 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 disease. PyRadiomics [63] widely adopted methods, which provide standardised feature definitions and extraction pipelines, enabling reproducibility and facilitating large-scale radiomics studies. With the support of AI, radiomics and radiogenomics are 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 [62-64].”

- Q3: Regarding the diffusion models for (biomedical) image restoration and

reconstruction, it doesn't not mean the denoising process in DDPM. The author is encouraged to search the recent works in diffusion model for inverse problem solving. There have been many recent works in past few years studying this area with significant application in medical imaging, for example, just to list two review papers [A, B]. The description in the revised version may be misleading.

[A] Daras, Giannis, et al. "A survey on diffusion models for inverse problems." arXiv preprint arXiv:2410.00083 (2024).

[B] Chung, Hyungjin, et al. "Diffusion models for inverse problems." arXiv preprint arXiv:2508.01975 (2025).

We thank the reviewer for this suggestion. We agree that diffusion-based inverse problem solving plays an important role in biomedical image restoration and reconstruction.

In the revised manuscript, we have updated the text to reflect the capabilities of diffusion models for solving inverse problems in biomedical imaging. We have also provided a brief discussion and cited the suggested survey papers. The revised text on page 7 is reported below.

“ 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].”

Reviewer #3 (Remarks to the Author):

The authors have done a great work in improving the paper aiming to show that the different works that have been exploiting multimodal learning within the medical imaging field.

Overall, I think the paper is in good shape and can be considered for publication. Nonetheless, I have one last relatively minor remark. In the challenges about data quality, the authors miss to mention the problem of missing modalities. This is a topic that has for long been disregarded, but that starts to gain traction as it is a critical problem that may hinder translation into practice, where a missing modality at inference can often occur. As pointers to some works that have started to look into this the authors may look into:

- https://doi.org/10.1007/978-3-031-72384-1_28

- https://doi.org/10.1007/978-3-031-84525-3_15

Please consider these (and perhaps others) in your discussion in section 4.1.

We thank the reviewer for this remark. We agree that missing data modalities present a significant challenge in multimodal biomedical imaging, particularly acquiring complete multimodal datasets is often costly and constrained by clinical factors such as limited scanning time, patient conditions, and machine availability. We appreciate the reviewer highlighting this important and timely issue.

In response to the reviewer's suggestion, we have expanded the challenge discussion in Section 4.1 (page 24) to address the problem of missing modalities. The revised text now highlights the AI-based solutions of ongoing research, such as generative AI and contrastive learning, for handling missing modalities in biomedical imaging.

The revised text is copied below:

“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 missing-modalities 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].”

#Reviewer 2 Report- Round 1

This review explores and discusses the recent artificial intelligence (AI) techniques applied in multimodal biomedical imaging, while also presenting key challenges and future directions. It outlines several integration strategies for combining diverse biomedical imaging data types and addresses challenges related to data quality, model interpretability, and ethical considerations. The manuscript is well-written and thoughtfully organized, featuring helpful illustrations—such as figures and tables—that effectively summarize relevant works and recent advances in the field of multimodal biomedical imaging.

While the paper presents interesting content and clear demonstrations, there are several areas where further clarification and refinement may enhance this review article.

- Section 1.1 provides a broad overview of various biomedical imaging modalities, including radiological imaging, microscopic imaging, histopathological imaging, and etc. This comprehensive summary is nice. However, the categorization of imaging modalities could benefit from clearer distinctions, as some categories appear to overlap or exist at different hierarchical levels. For instance, ultrasound imaging is a subset of radiological imaging; microscopic imaging can include both histological and histopathological techniques; and modalities such as multiplexed or data-driven imaging may refer more towards the image processing and analytical approaches rather than to distinct imaging types. It may be worthwhile to consider reorganizing the categories to better reflect their functional or hierarchical relationships and improve conceptual clarity.
- Section 2 raises similar concerns regarding the logical structure of the discussion. From an image processing perspective, techniques such as normalization and denoising are typically components of a broader category of image enhancement, which may also include deblurring, super-resolution and etc., depending on the imaging modality and task requirements. Additionally, while feature extraction is often associated with convolutional neural networks (CNNs) in modern contexts, it may not be limited in this scope. Traditional hand-crafted methods such as Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT) remain relevant and effective. Furthermore, tasks like image segmentation are more accurately described as distinct image analysis objectives, similar to classification, detection, and localization. Clarifying these distinctions could help improve the logical flow of the section.
- In Section 2.1, the use of generative models such as GANs, VAEs, and diffusion models for data augmentation is mentioned. While generative models can indeed assist with data augmentation, it is worth noting that their effectiveness in improving downstream predictive performance remains an area of active research with existing debating on this topic. Therefore, it may be advisable to present a more balanced view by noting these potential issues and discussions, and avoid labeling them only as data augmentation approaches. Instead, these generative models have shown to be very useful in image reconstruction and restoration tasks in biomedical imaging applications, in the cutting-edge research.
- Section 3.1 presents an overview of multimodal image fusion strategies categorized as

pixel-level, feature-level, decision-level, and hierarchical-level approaches. While this framework is informative, it is important to note that prior literature often refers to these strategies as early, intermediate, late, and joint fusion, respectively, as discussed in [1]. Connecting these terms to existing literature would strengthen the contextual grounding of the discussion.

- Table 2 offers a helpful summary of different previous works based on various integration techniques. To further enhance the review, it may also be beneficial to include an additional table categorizing existing works based on specific tasks or anatomical sites. This could provide a clearer benchmark of how multimodal images, or different approaches address particular clinical or analytical challenges.
- Regarding the discussion of multimodal large language models (MLLMs) as a prospective future direction, it is worth highlighting that recent studies have already begun to demonstrate their potential in biomedical imaging contexts. A more detailed review of these emerging works, possibly integrated into the main text rather than limited to a brief mention in the future directions, would add great value to make this review article more interesting and timely.
- Similarly, Section 5.3, which discusses the integration of imaging data with other data modalities, touches on a highly active area of current research. Expanding this section to include a more thorough survey of recent advancements would likely resonate with readers interested in this growing subfield.
- Finally, while the review offers a broad and informative overview, it would be helpful for the authors to articulate more clearly what distinguishes this work from existing review articles on multimodal biomedical imaging. Clarifying the novel insights or unique perspectives provided in this paper could better highlight its contribution to the field.

[1] Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, npj Digital Medicine, 2020.

