Navigating the landscape of multimodal AI in medicine: a scoping review on technical challenges and clinical applications

Daan Schouten[a,g,1,*], Giulia Nicoletti[b,c,1], Bas Dille[d,a,1], Catherine Chia[a,e,f], Pierpaolo Vendittelli[a], Megan Schuurmans[a], Geert Litjens[a,g], Nadieh Khalili[a]

[a]*Department of Pathology, Research Institute for Medical Innovation, Radboud University Medical Center, Nijmegen, the Netherlands*
[b]*Department of Electronics and Telecommunications, Polytechnic University of Turin, Turin, Italy* [c]*Department of Surgical-Medical Sciences, Unversity of Turin, Turin, Italy*
[d]*Department of Radiology and Nuclear Medicine, Erasmus University Medical Centre, Rotterdam, the Netherlands*
[e]*Department of Dermatology, Erasmus University Medical Centre, Rotterdam, the Netherlands*
[f]*Department of Pathology, Erasmus University Medical Centre, Rotterdam, the Netherlands*
[g]*Oncode Institute, Utrecht, the Netherlands*

Abstract

Recent technological advances in healthcare have led to unprecedented growth in patient data quantity and diversity. While artificial intelligence (AI) models have shown promising results in analyzing individual data modalities, there is increasing recognition that models integrating multiple complementary data sources, so-called multimodal AI, could enhance clinical decision-making. This scoping review examines the landscape of deep learning-based multimodal AI applications across the medical domain, analyzing 432 papers published between 2018 and 2024. We provide an extensive overview of multimodal AI development across different medical disciplines, examining various architectural approaches, fusion strategies, and common application areas. Our analysis reveals that multimodal AI models consistently outperform their unimodal counterparts, with an average improvement of 6.2 percentage points in AUC. However, several challenges persist, including cross-departmental coordination, heterogeneous data characteristics, and incomplete datasets. We critically assess the technical and practical challenges in developing multimodal AI systems and discuss potential strategies for their clinical implementation, including a brief overview of commercially available multimodal AI models for clinical decision-making. Additionally, we identify key factors driving multimodal AI development and propose recommendations to accelerate the field's maturation. This review provides researchers and clinicians with a thorough understanding of the current state, challenges, and future directions of multimodal AI in medicine.

Keywords: Multimodal AI, Deep Learning, Multimodal data integration, Medical Imaging

1. Introduction

The healthcare landscape is evolving rapidly, driven by an increasingly data-centric approach to patient care and decision-making (Shilo et al., 2020). This shift is complemented by the advent of technologies such as digital pathology (Niazi et al., 2019), biosensors (Sempionatto et al., 2022), and next-generation sequencing (Steyaert et al., 2023), which provide clinicians with novel insights in various domains. The data generated by these diverse modalities is generally complementary, with each modality contributing unique information to the status of a patient. Some modalities offer a comprehensive overview at the macro level, while others may provide detailed information at single-cell

* Corresponding author. (*email:* Daan.Schouten@radboudumc.nl)
[1]These authors contributed equally to this work

resolution (Steyaert et al., 2023). In addition to this recent growth in data quantity, there is a concurrent increase in the quality and diversity of available treatment options. Hence, selecting the optimal treatment has become increasingly complex, and a further data-centric approach to treatment selection may be required.

The traditional approach to integrating information from different data modalities into a single decision is represented by multidisciplinary boards, where each specialized clinician offers their perspective on a given modality or piece of information in pursuit of consensus (Mano et al., 2022). Although establishing these boards has improved disease assessments and patient management plans (Mano et al., 2022), there is a foreseeable limit to the scalability of these boards. If data quantity and diversity continue to rise, many domain experts will be required to integrate these different information streams effectively. Fortunately, another technological advancement that is gaining a foothold in healthcare is artificial intelligence (AI). Although the vast majority of published work focuses on single modality applications of AI, several authors have highlighted the potential of AI systems to combine multiple streams of information, so-called multimodal AI, for decision-making (Steyaert et al., 2023; Acosta et al., 2022; Lipkova et al., 2022). These multimodal AI models are trained to process different streams of multimodal data effectively, leverage the complementary nature of information, and make an informed prediction based on a broader context of the patient's status. However, despite these promising results, studies investigating multimodal AI models are comparatively scarce, and the development of unimodal models remains the de facto standard.

This lagging development of multimodal AI models can be attributed to several challenges. First, a practical challenge can be found in the cross-departmental nature of multimodal AI development. As different data modalities may originate from various medical departments, consulting different medical domain experts will likely be required for effective data integration. In addition, medical departments may have varying experience in data storage, retrieval, and processing, limiting the possibilities of multimodal AI development. For example, if a radiology department has a fully digital workflow while the corresponding pathology department does not, this effectively prohibits multimodal AI endeavors where whole slide images would be combined with radiological imaging data.

Different data modalities can have vastly different characteristics, such as dimensionality or color space, which generally requires different AI model architectures tailored towards those modalities, increasing model design complexity. For example, convolutional neural networks (CNN) were initially proposed for structured data, such as 2D and 3D images, but can't straightforwardly be applied to unstructured data. Conversely, transformers are solid, flexible encoders for various data modalities. Still, whether a one-size-fits-all architecture can capture various medical data modalities effectively remains unclear. In practice, multimodal data integration is commonly achieved using different (intermediate) model outputs. Training multiple domain-specific AI models (i.e., encoders) and efficiently integrating these in a single prediction poses a challenge unique to multimodal AI development.

Last, the inconsistent availability of all modalities for each patient within a multimodal dataset adds complexity. Patients with different disease trajectories will have various available modalities, leading to partially incomplete datasets. This can substantially reduce the adequate training dataset size for AI models that require complete multimodal data to generate predictions. Moreover, these issues also translate to implementation. If modalities are missing, it may be unclear how this impacts the model's performance from the perspective of fewer available data to base a decision on and the potential introduction of population selection bias (Acosta et al., 2022). In short, developing multimodal AI models poses several novel challenges compared to unimodal AI development.

Even given these challenges, several works have been done in the past on multimodal AI applications, typically involving handcrafted features. A key issue with these approaches was that the difficulties requiring particular domain expertise are multiplied, as expert clinicians

would also need to be involved in the feature design phase (Vaidya et al., 2020; Tortora et al., 2023). An excellent overview was published by Kline et al. (2022), indicating that these models obtained a 6.4% mean improvement in AUC compared to their unimodal counterparts.

Recent years have shown an accelerated interest in multimodal AI development for medical tasks (Salvi et al., 2024), as using unsupervised learning and deep neural networks as encoders has significantly simplified the feature extraction step. In this review, we comprehensively summarize the state-of-the-art in multimodal AI development for medical tasks and investigate to what extent multimodal data integration is living up to its purported benefits. Unlike previous reviews that have focused on specific diseases, prediction tasks, or modality combinations (Acosta et al., 2022; Salvi et al., 2024; Krones et al., 2025), our analysis encompasses the full spectrum of the medical domain. Specifically, our review aims to shed light on I) the progress of multimodal AI model development across different medical disciplines and tasks, II) the technical challenges inherent in multimodal AI development, including model architectures, fusion methods, and the handling of missing data, III) the foreseeable road to the clinic of multimodal AI models, addressing aspects such as regulatory approval and explainability, and IV) the factors driving multimodal AI development and potential strategies to promote further maturation of this field. Lastly, we will provide an outlook on future perspectives in multimodal AI development based on our careful analysis of 432 papers published over the past six years (2018-2024).

2. Search Criteria

This scoping review aimed to evaluate the application of multimodal AI models in the medical field, where we define multimodality as data originating from different medical specialties. For instance, we consider a model to be multimodal when it integrates a diagnostic CT scan and subsequent tissue biopsy slides, as it combines the domains of the radiologist and the pathologist, respectively. Conversely, a model integrating T1and T2-weighted MRI scans would not be considered multimodal. In addition to this multimodality criterion, we limited our scope to I) studies using deep neural networks and II) studies developing multimodal models for specific medical tasks (i.e., no generic visual question answering).

The literature search was conducted in PubMed, Web of Science, Cochrane, and Embase. The entire search string per database can be found in the supplementary materials. The search was initially performed on April 16, 2024 and repeated on October 2, 2024 to ensure an up-to-date overview, yielding a total of 12856 initial results. These results included both journal and conference papers. In addition to the previously mentioned inclusion criteria, we applied additional standard exclusion criteria summarized in Figure 1. Specifically, we excluded review articles, non-English publications, articles without full text, non-peer-reviewed preprints, and those published before 2018. After filtering with these criteria and deduplicating results using the DedupEndNote tool (Lobbestael, 2023), 10522 articles were imported into the paper screening software Rayyan (Ouzzani et al., 2016) for the Title and Abstract (TiAb) screening phase. A team of six reviewers conducted the TiAb screening phase. Each paper's title and abstract were reviewed based on the inclusion criteria. Included papers were verified by a second reviewer, and potential discrepancies were resolved through discussion or the involvement of a third reviewer if necessary. The TiAb screening phase was performed sequentially, starting with verifying that the paper is in the medical domain, then determining whether it is in scope, whether multimodality is involved, and finally, assessing whether deep neural networks are used. Papers were deemed to be outside the scope of this review when they investigated tasks that do not have an explicit clinical question (i.e., image denoising, registration). The TiAb screening resulted in 663 articles being considered for full-text screening.

In the full-text reading phase, a single reviewer read each article in full to confirm adherence to our inclusion criteria. In case of doubt, a second reviewer was involved

to arrive at a consensus decision. Eventually, 432 studies were included in the final analysis (see the supplementary materials for the entire list of included papers).

3. Overview of multimodal medical AI

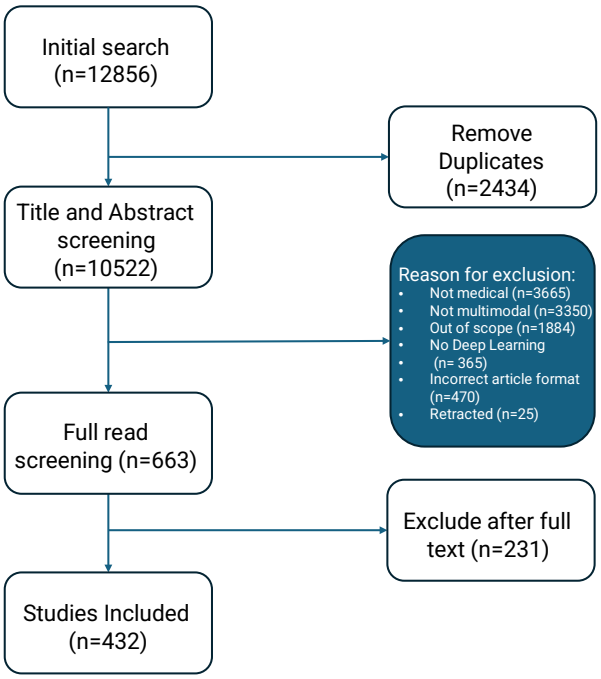
The landscape of multimodal artificial intelligence (AI) in medical research has expanded between 2018 and 2024, as shown by the growing number of articles focused on integrating multiple modalities. Subdividing the 432 articles in this review per year (Figure 2a), we see a rapid increase, starting with 3 papers in 2018 to 150 papers in 2024 at the end of the data collection of this review.

We distributed the reviewed papers according to data modalities, which will be broadly described here.

Subsequently, the review is organized according to the following subsections:

- State-of-the-art algorithm designs for multimodal medical AI (section 4)
- Clinical tasks (section 5) for a detailed analysis of the multimodal AI approaches categorized per organ system
- Challenges and opportunities for clinical adoption (section 6) of multimodal AI

Last, we will conclude with a high-level discussion of our findings and provide guidelines for future research directions.



3.1. Modalities and data types

We divided the data modalities into image-based and non-image-based modalities. The image-based modalities are grouped according to their related medical specialties. This resulted in the following categories: radiology (computed tomography (CT), magnetic resonance imaging (MRI), ultrasound (US), X-rays, and nuclear imaging (SPECT/PET)), pathology (stained histology images) and 'clinical images' (including optical coherence tomography (OCT), fundus photography, and dermatoscopy, among others). If there were too few papers for a medical specialty, we grouped them into a catch-all 'other images', a subset of 'clinical image' category. Meanwhile, the non-image-based modalities consist of text (structured text such as tabular laboratory results and unstructured text such as free-text medical reports), omics data (e.g., genomics, transcriptomics, or proteomics), and other non-image modalities (such as

Figure 1: Overview of the screening process.

showcase a preference for integrating one imaging modality with structured or unstructured text data. A complete overview is presented in Figure 2d.

The review also reveals several complex combinations involving three or more modalities, albeit in smaller numbers. Notably, pathology/text/omics (n=19), radiology/text/omics (n=15), radiology/pathology/text (n=7), and radiology/pathology/omics/text (n=3) demonstrate the attempts to create comprehensive models that span multiple scales of biological organization - from organ-level (radiology) to tissue and cellular (pathology) and subcellular (omics), complemented by clinical context (text). An exhaustive overview of all modality combinations can be found in the supplementary materials.

3.2. Organ systems, medical tasks and AI functions

We categorized all studies into eleven organ systems (see Figure 3a for the complete list). The nervous system dominates with 122 studies, followed by respiratory (n=93), reproductive (n=43), digestive (n=43), sensory (n=25) and integumentary (n=24). A substantial number of studies (n=15) fall under the miscellaneous category and 27 studies involved multiple organ systems, which are further explained in the dedicated section Clinical applications (section 5).

Six medical task categories: diagnosis, estimating prognosis (which is further subdivided into survival prediction, disease progression, and treatment response analyses), treatment, and others. Diagnosis emerged as the primary focus, accounting for 45% (multiple systems) to 91% (integumentary system) of the medical tasks across all organ systems (see Figure 3a and b). Survival prediction was the second most common task (18% of all medical tasks). Other tasks, such as prediction of disease progression or treatment response, were not uniformly represented across all organ systems.

Electroencephalography (EEG) or Electrocardiography (ECG) signals).

Radiology and text were the most commonly used modalities (each 30%), followed by omics (12%) and pathology (12%). Figure 2b visualizes all modalities and their respective subtypes. Over the years, the trend of using radiology and text modalities remained the most common (see Figure 2c). The most prevalent combination is radiology with text (n=206), followed by pathology/omics (n=51), clinical images/text (n=33), radiology/omics (n=24), pathology/text (n=22), and radiology/pathology (n=16). These combinations

4. Methodology

4.1. Importance of public data

As stated in the introduction, data availability is a key challenge for the development of multimodal medical AI. This is why we see a strong correlation between the number of models for a specific organ system/modality combination and the availability of public data (see Figure 3c). The utilization of publicly shared datasets in multimodal AI research for medical applications is widespread, with 61% of the data sources used in the model development coming from public data portals such as The Cancer Genome Atlas (TCGA, 14%), Alzheimer's Disease Neuroimaging Initiative (ADNI, 8%), Medical Information Mart for Intensive Care (MIMIC, 5%) and The Cancer Imaging Archive (TCIA, 2%), 15% from data shared publicly through other means (e.g. GitHub, publisher's website), and 24% from private datasets which were not shared publicly. We grouped all other data portals used by less than ten reviewed papers into "other data portals" (20%). A detailed breakdown of these public data sources can be viewed in the supplementary materials.

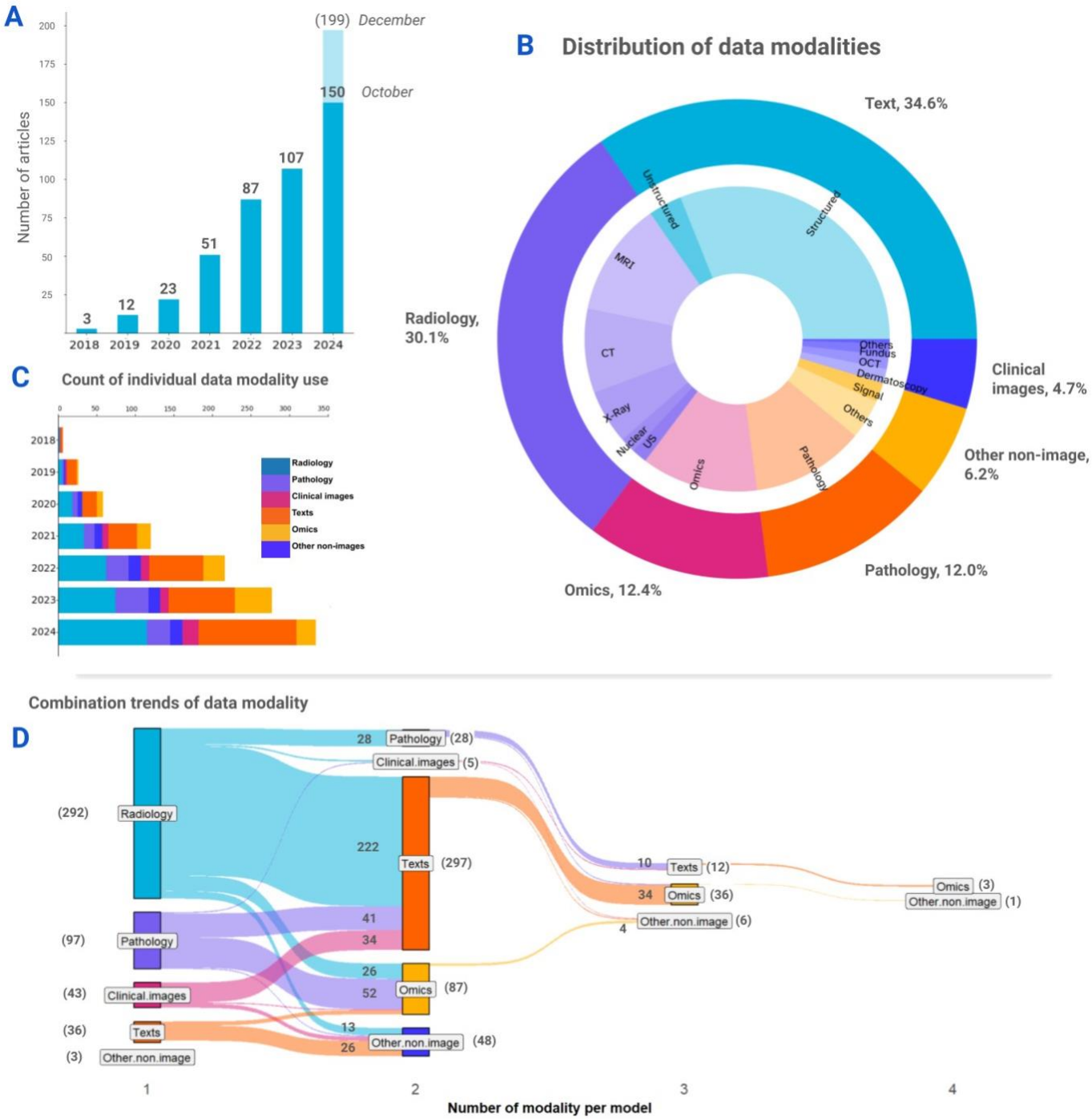
4.2. Feature encoding and modality fusion

The advent of deep neural networks was a key accelerator for multimodal medical AI. These networks significantly simplified the feature extraction/encoding step for individual modalities. Each modality could now be encoded by its deep neural network, and the resultant features could be combined for downstream tasks. Powerful self-supervised learning techniques such as DINO (Caron et al., 2021) and SimCLR (Chen et al., 2020) have now also enabled the training of these feature encoders without any labels, further increasing the attractiveness of this approach. However, we have seen that there is still significant diversity in approaches across

the reviewed papers, which we will detail in the following subsections.

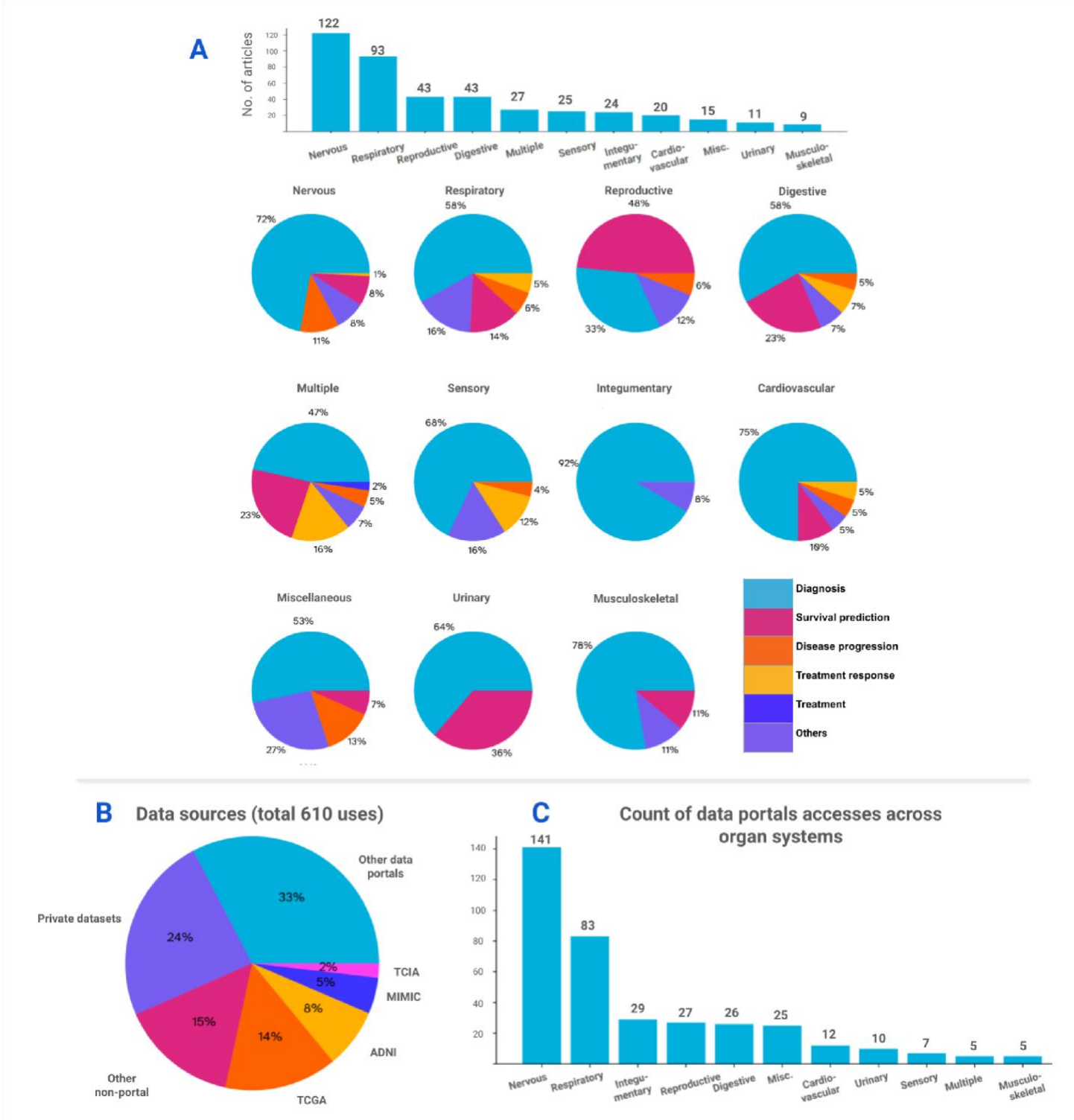
Various encoder mechanisms are used for different data modalities. Encoders are categorized into Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), Recurrent Neural Networks (RNNs), Transformers, handcrafted feature encoders, multi-layer perceptrons (MLPs), multiple encoders, and 'other' (e.g. graph neural networks). Our analysis reveals that CNNs dominate the encoder landscape (used in 82% of the studies), followed by 'other' (32%), and MLPs (21%). Unsurprisingly, CNNs show strong correlations with image-based data modalities, including radiology, pathology, and other (miscellaneous) image modalities. In contrast, non-image modalities, such as 'omics and (un)structured text, use

Figure 2: Overview of the data modalities used in the reviewed articles. (A) Distribution of articles by year. Bar chart shows an exponential increment in the number of studies per year from 2018 to 2024. Extrapolating, the number of multimodal medical AI studies is expected to reach 199 by the end of 2024. (B) Pie chart shows the proportions of different modality groups and the respective data modalities used across studies. (C)6 Stacked bar chart illustrates the growth trends of data modality groups over the years. Note that the values used in this chart represent the counts of individual data modality uses, where multiple modalities could be presented in a single article. (D) Diagram shows the combination



trends between data modalities per model. The diagram captures the unique modality combinations presented in each models the individual article has presented. The numbers in brackets indicate the total summation of models per category, whereas the numbers without brackets represent the count of models of each combination, visualized with the ribbon bands between the vertical nodes. The majority of the models used two data modalities, and a portion of the total used three and four modalities. Three multimodal models used data modalities that were grouped under ”other non-image” category based on the definition used in this review.

Figure 3: A deeper dive into the medical tasks and data sources of the review. The numbers on the bars indicate the total summation per category. (A) Top: The number of articles per organ system. Bottom: Distribution of medical tasks across organ systems. Pie charts show diagnosis being the most prevalent medical task performed in studies of all organ systems. (B) The use trends of data sources in this review. Note that the values used in the chart represent the total count of uses of all the reviewed studies, where multiple data sources could be



referred to in each study. About 61% of the total uses were sourced from data portals (e.g. TCGA, ADNI, etc.), 15% from research data shared publicly by publications, and 24% of the data uses were private datasets that were not made public. (C) Distribution of public data sources

(excluding private datasets) across the studies of organ systems. Similarly, the nervous and respiratory systems are leading in the count of public data uses. A detailed breakdown of these public data sources can be found in the supplementary materials.

a more diverse range of encoders, including handcrafted feature encoders, MLPs, RNNs, and Transformers.

The second important design decision is how and when the differing modalities are fused. The fusion stage can vary depending on the model structure, which often dictates the optimal point for integrating information across modalities. A schematic view of these fusion stages is shown in Figure 4.

Out of all papers reviewed in this study, the vast majority (79%, 341/432 papers) utilized intermediate fusion, in which data sources get fused after feature encoding but before the final layers of the neural network (e.g., the classification or regression head). A common strategy is simply concatenating the feature vectors of the different unimodal modality encoders and feeding the resultant vector to the final layers. This concatenation method was used in most models (69%) using intermediate fusion, and was not limited to certain modality combinations or clinical application areas (Lee et al., 2019; Zhi et al., 2022; Liu et al., 2024). However, several studies found through ablation experiments that applying other methods to fuse the feature vectors (i.e. taking the outer product, Kronecker product or compact bilinear pooling), outperformed concatenation by a notable margin (Yang et al., 2022; Wang et al., 2023b, 2024a). Another common intermediate fusion technique (12%) was the use of attention, where the unimodal embeddings were passed through an attention mechanism to optimally learn from the complementary information in both embeddings (Kayikci and Khoshgoftaar, 2023; Liu et al., 2023; Machado Reyes et al., 2024).

Late fusion, a method in which fusion is performed by combining results or predictions of unimodal models, is the second most common technique (14%). These architectures can look similar to those from intermediate fusion models, but without intermediate components such as attention mechanisms or mixing layers. However, most of these late fusion approaches combine unimodal models and their individual predictions to get a multimodal result. This was often (32%) achieved by applying a (weighted) average over the predictions for each modality (Ying et al., 2021; Caruso et al., 2022; Jung et al., 2024)) or by training a separate model on top of the unimodal predictions (37%). The latter was mainly done through regression or traditional machine learning models like random forests, boosting algorithms or Cox models (Ma and Jia, 2020; Kolk et al., 2024; Wang et al., 2024b). A key aspect of late fusion is that no interaction exists between the modalities in the learning process, meaning that the training data for the respective unimodal models does not have to be paired. It therefore becomes easier to handle missing data in these models, as patients with only one modality can be used to train only the model of that modality without the need to infer the missing ones. On the contrary, the lack of interaction could potentially limit model expressiveness.

The last method involves early fusion, at which modalities are fused before feature encoding. Our study found that this fusion method has been applied the least (6%). A key challenge is that the input data has to exist in the same 'space' to allow early fusion. The difficulty beyond this kind of fusion could be influenced by the types of modalities involved in the fusion, for example, combining radiology and pathology images might often require some form of image registration, which is an unsolved challenge for many clinical applications. However, several early fusion methods were implemented that do not require extensive preprocessing steps, ranging from a simple concatenation of the modalities (Shi et al., 2023; Lopez et al., 2020) to more complex representations involving graph networks (Lei et al., 2024), recurrent neural networks (Xu et al., 2022a), and cross-modality representation alignment networks (Wu et al., 2023). Other interesting methods combined image and non-image modalities by directly marking (Pelka et al., 2020) or multiplying (Qiu et al., 2024) clinical variables onto images, thereby enhancing the image data with relevant contextual information. A key potential advantage of early fusion is that modalities are already available in the feature encoding stage, potentially allowing deep neural networks to optimize feature design by leveraging all information simultaneously.

4.3. Network architectures

The architecture choices for multimodal AI models often rely on the purpose(s) of the model and data availability. Regardless, the critical steps when designing such a model include the same core functionalities in feature extraction, information fusion, and final processing of the fused information. Some papers also introduce explainability (Ketabi et al., 2023; Yang et al., 2021; Parvin et al., 2024) into their models or attempt to

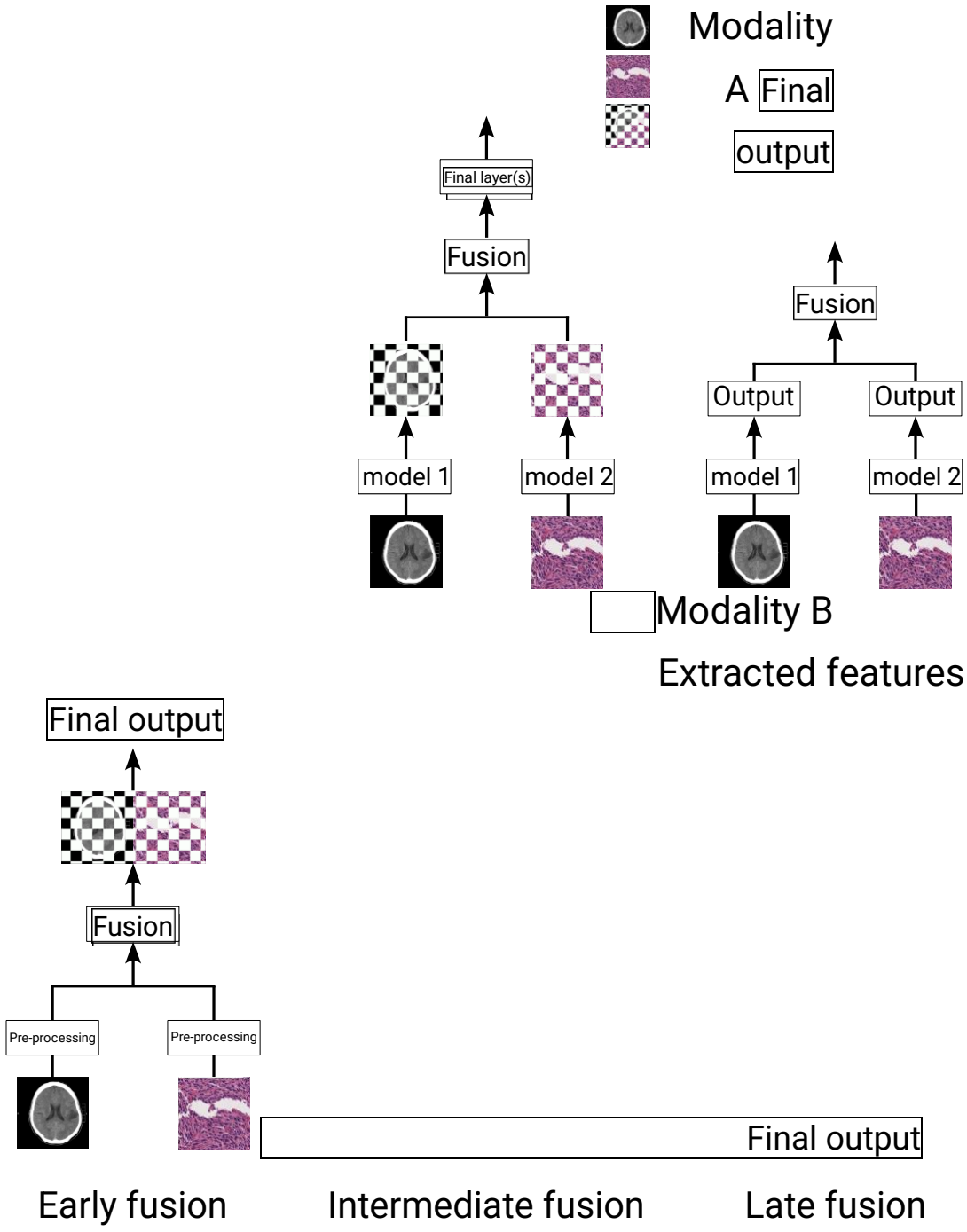


Figure 4: Simplified schematic view of the different fusion stages.

construct multiple different model architectures to assess the effectiveness of each variant (Huang et al., 2020; Lopez et al., 2020; Caruso et al., 2022). Examining the usage of individual modality encoders over the years reveals that contemporary feature extraction models like (vision) transformers are gaining popularity, while on the other hand, MLPs have seen a decrease in usage, with CNN usage staying roughly the same in terms of popularity over the last few years. Traditional machine learning methods are also still employed for feature encoding in 5% of all papers, often to extract features from structured text data.

4.4. Handling missing modalities

Most multimodal AI models work on the assumption of data completeness: all modalities must be available for every entry. However, this is often not feasible due to issues such as data silos caused by incompatible data archival systems (i.e. PACS), a study's retrospective nature, or simply data privacy concerns. Handling missing data without introducing bias into the analysis presents a significant challenge. Indeed, a commonly employed approach to address this issue is to exclude all entries with at least one missing modality and include only complete entries in the dataset used for analysis, and this is the case for 69% of the included papers. Although the exclusion of incomplete entries is a quick and straightforward method for handling missing data, this approach results in a reduction in data sample size with a subsequent loss of potentially relevant samples. Furthermore, this limits the model inference to complete-modality data and inevitably leads to selection biases, thereby potentially compromising validity and generalizability. Rather than discarding incomplete samples, missing information can also be estimated using imputation techniques, thereby alleviating both aforementioned disadvantages of fully excluding incomplete entries. These data imputation techniques can be broadly divided into non-learning-based and learning-based approaches.

4.4.1. Non-learning-based approaches

The findings of our study indicate that non-learning-based approaches are commonly utilized, occurring in 45% (35/78) of all papers that reported any method to account for missing data. These techniques are primarily employed to impute values of structured data (n=29), such as clinical variables and test results. Less frequently, these methods are used to address missing gene values (n=2) or image modalities (n=2). For continuous variables, imputation is performed using measures of central tendency such as the mean or median, or performing a moving average, which replaces missing values with the average of a specified number of surrounding data points. For categorical variables, imputation often involves using the mode or adding a new category for missing values. Some studies apply fixed values, like zeros, -1, or random values from similar records (hot-deck method). Although these methods are simple and easy to implement, they can introduce bias and reduce data variability (Flores et al., 2023).

4.4.2. Learning-based approaches

Some studies leveraged traditional machine learning methods to predict missing values, learning patterns from complete data as an alternative to simpler imputation techniques. Common methods included k-nearest neighbor (k-NN), which imputes missing values based on the values of the nearest neighbors (Ross et al., 2024; Qiu et al., 2022; Lee et al., 2024b), and the weighted nearest neighbor approach (Wu et al., 2023; Kayikci and Khoshgoftaar, 2023; Mustafa et al., 2023; Palmal et al., 2024), which assigns weights to neighbors based on their distance. Other techniques used include linear regressions (Ghafoori et al., 2023), random forests (Yu et al., 2024; Kolk et al., 2024), neural networks (Menegotto et al., 2021) and Classification and Regression Trees (CART) algorithm (Yin et al., 2024), as well as advanced methods like Multivariate Imputation by Chained Equations (MICE) algorithm (Rahman et al., 2023; Liu et al., 2022; Lim et al., 2022; Kim et al., 2024) and XGBoost (Feher et al., 2024; Zambrano Chaves et al., 2023; Fan et al., 2024), which handle missing data by respectively incorporating multiple imputations and tree-based learning during training.

Other methods applied deep learning models to impute missing values or manage an arbitrary number of modalities, directly modifying the model architecture to process incomplete data without imputation. Some deep-learning-based imputation strategies leverage different generative models, such as auto-encoders (Xu et al., 2022b; Akramifard et al., 2021), or generative adversarial networks (Dolci et al., 2023). Others directly predict the missing data at the output layer (Saad et al., 2022) or at previous visits of a Recurrent Neural Network (Xu et al., 2022a). However, there seems to be little consensus or evidence to suggest that one method should be preferred. Some papers have addressed the

missing data challenge by introducing specific drop-out modules to train on or simulate missing data (Cheerla and Gevaert, 2019; Ostertag et al., 2023; Cui et al., 2022b; Liu et al., 2023). Other studies addressed the missing modality problem by designing specific loss functions that take into account only available modalities (Gao et al., 2021; Xue et al., 2024; Kawahara et al., 2019; Nguyen et al., 2024; Taleb et al., 2022), sometimes employing a reconstruction loss to reconstruct them (Cui et al., 2022a). Interestingly, some studies used learnable embeddings as placeholders for missing modalities (Chen et al., 2024) or directly utilized models, such as transformers, capable of handling input sequences of arbitrary lengths (Zhou et al., 2024b), allowing the model to manage cases with missing or incomplete modalities effectively.

4.5. Validation

An observation on the validation of multimodal AI models is that most studies (82%) are limited to an internal validation scheme. Despite the common use of public datasets, these are often employed as the sole training data rather than as external validation. Another important component in validating multimodal models is the comparison with a unimodal baseline. To provide a broadly representative analysis of the performance gains by multimodal data integration, we equally sampled 3-4 papers per organ system, resulting in a subset of 48 papers that clearly described a multimodal vs unimodal baseline comparison. On average, these multimodal models obtained a 6.2 percentage point increase in AUC, which aligns with the 6.4 percentage points improvement reported previously by Kline et al. (2022). From the 432 papers studied, 72% of the papers mention an improvement over unimodal models in an ablation experiment. In contrast, only 5% of the papers did not find any notable improvement by including multiple modalities, and 22% of papers did not report any comparison with a unimodal baseline. Despite these encouraging findings, it should be noted that these improvements are rarely tested for significance, leading to a potentially optimistic bias.

5. Clinical applications

Subdividing multimodal AI research papers based on the applied organ system reveals significant variations in the development of multimodal models. The distribution of the papers among the systems is shown in Figure 3a. Interestingly, some studies evaluated their models into more than one system and, therefore, are reported in a separate *Multiple Systems* category (n=27). Papers that did not align with any category were grouped into a *Miscellaneous* category (n=15). Below, we summarize the key contributions in each area.

5.1. Nervous system (n=122)

The predominant focus in the nervous system has been on the diagnosis and disease progression of neurodegenerative disorders, such as Alzheimer's disease (n=47), with a few studies focusing on Parkinson's disease (n=9). The second primary focus is cancer diagnosis and survival prediction (n=22). Other studies focused on the diagnosis of autism (n=6), stroke (n=9), and mental disorders, such as schizophrenia (n=6) and depression (n=4). The analyses of these papers showed that most studies in this area utilize MRI as the primary modality, integrating either clinical data (n=43) or 'omics data (n=18), or all three (n=8). Other studies combined pathology with MRI (n=6) or 'omics (n=5), or CT with clinical variables (n=10).

A significant number of studies utilized publicly available datasets (n=115), with the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset being the most frequently used (n=44), followed by The Cancer Genome Atlas (TCGA)(n=11), the 'Computational Precision Medicine Radiology-Pathology challenge on brain tumor classification' dataset (CPM-RadPath) (n=6) and Brain Tumor Segmentation Challenge (BraTS) (n=3). Only a limited number of studies conducted external validation of their models (n=15), while the majority employed cross-validation (n=63) and internal validation (n=40). Integrating multiple modalities has enhanced performance in most of these studies (n=83), underscoring the value of incorporating additional data

sources that capture information not evident in a single modality alone.

Some studies stood out for their system design and development, with several focusing on glioma grading, such as Li and Ogino (2020), which proposed a patient-wise feature transfer model that learns the relationship between radiological and pathological images. Notably, their model enables inference using only radiology images while linking the prediction outcomes directly to specific pathological phenotypes. Another interesting example is Cui et al. (2022a), which integrates histological images and genomic data and can handle patients with partial modalities by leveraging a reconstruction loss computed on the available modalities. Interestingly, Cui et al. (2022b) combined four different modalities, i.e., radiology images, pathology images, genomic, and demographic data, to predict glioma tumor survival and reached a c-index of 0.77 on a 15-fold Monte Carlo cross-validation.

Several studies focused on Alzheimer's prediction, including Liu et al. (2023), that is the only study both handling missing data and externally validating its findings, employing multiple transformer architectures to integrate imaging data and clinical variables and reaching an AUC value of 0.984 on an external validation composed of 382 patients. Pelka et al. (2020) proposed a unique type of early fusion where sociodemographic data and genetic data are branded on the MRI scan and used to develop a model able to diagnose early and subclinical stages of Alzheimer's Disease. Lei et al. (2024) developed a novel framework for AD diagnosis that fuses four modalities: genomic, imaging, proteomic, and clinical data, validating the proposed method on the ADNI dataset and outperforming other state-of-the-art multimodal fusion methods.

The most extensive study of this category is the one of Xue et al. (2024) that presented an AI model that integrates diverse data, including clinical data, functional evaluations, and multimodal neuroimaging, to identify 10 distinct dementia etiologies, across 51,269 participants from 9 independent datasets, even in the presence of incomplete data.

5.2. Respiratory system (n=93)

Research in the respiratory system predominantly focused on diagnosis (n=10), survival (n=9), treatment response prediction (n=3), and progression (n=2) of cancer, and on diagnosis (n=19), survival (n=3) and disease progression (n=2) of Covid-19. Across these studies, the common strategy involves combining clinical variables with either X-ray or CT imaging of the thorax (n=18 and n=27, respectively). Additionally, a subset of papers (n=11) developed vision-language models by integrating chest X-rays with clinical reports, specifically for X-ray report generation. The majority of the studies regarding respiratory diseases employed public datasets, such as the MIMIC-CXR dataset (n=9), The Cancer Genome Atlas (TCGA) (n=5), and the National Lung Screening Trial (NLST) dataset (n=3). In this category, many studies performed better than single-modality models when performing multimodal integration (n=69). However, the proportion of studies employing external validation is still limited (n=15).

Among the studies of the respiratory category that stood out, Keicher et al. (2023) proposed a multimodal graph-based approach combining imaging and non-imaging information to predict COVID-19 patient outcomes. Thanks to the employed attention mechanism, the model learns to identify the neighbors in the graph that are the most relevant for the prediction task, providing insight into the decision process. Gao et al. (2022) outperformed state-of-the-art models on three different external validation sets in predicting the risk of indeterminate pulmonary nodules with a multimodal approach combining CT imaging and clinical variables, where most studies do not perform external validation. Gao et al. (2021) proposed a 'multi-path multimodal missing' network, a system integrating multimodal data, including clinical data, biomarkers, and CT images, that can be trained end-to-end with even incomplete data types. The resultant model can make predictions using even a single modality. Cross-validation and external validation showed that combining multiple modalities significantly improves performance in predicting lung cancer risk compared to using a single modality. Wang et al. (2024c) proposed a novel lung cancer survival analysis

framework using multi-task learning, incorporating histopathology images and clinical information, reaching a cross-validation C-index of 0.73 that outperformed traditional methods. Kumar et al. (2023) combined X-ray with clinical information to develop a multimodal fusion approach to detect lung disease and developed two multimodal network architectures based on late and intermediate fusion, showing better performances with the latter. Lopez et al. (2020) examined all three multimodal fusion types, early-intermediate-late, with deep learning-based techniques for classifying several chest diseases using radiological images and associated text reports, demonstrating the potential of multimodal fusion methods to yield competitive results using less training data than their unimodal counterparts.

5.3. Digestive system (n=43)

In the domain of the digestive system, most studies focused on the diagnosis of malignancies (n=25), with fewer addressing survival (n=10), treatment response (n=3), and disease progression prediction (n=2). The primary malignancies investigated include colorectal (n=12) and liver (n=10) cancers, with some studies also focusing on the stomach, esophagus, and duodenum (n=9). No modality combination was dominant in this field, though clinical variables (n=31) and histopathology slides (n=17) were commonly included. Integrating multimodal data improved performance in 38/39 papers compared to unimodal approaches. Despite being the category with the highest proportion of externally validated studies (n=11), external validation remains limited. Although quite some studies employed publicly available datasets (n=16), including mainly TCGA (n=11), these datasets were often used for training rather than external validation.

Several studies were particularly noteworthy. Cui et al. (2024) developed a multimodal AI model using both endoscopic ultrasonography images and clinical information to distinguish carcinoma from noncancerous lesions of the pancreas and tested this model in internal, external, and prospective datasets. They also evaluated the assisting potential of the model in a crossover trial. Their results showed that AI assistance significantly improved the diagnostic accuracy of novice endoscopists and that the additional interpretability information helped reduce skepticism among experienced endoscopists. Chen et al. (2024) introduced a deep learning model combining three modalities, radiology, pathology, and clinical data, to predict treatment responses to anti-HER2 therapy or anti-HER2 combined immunotherapy in patients with HER2-positive gastric cancer. This model can manage missing modalities thanks to the incorporation of learnable embeddings which, as substitutes for the modality-specific features, are concatenated to the features of the existing modalities to generate the inter-modal fused feature. Zhou et al. (2023) developed and externally validated a multimodal model to predict treatment response to bevacizumab in liver metastasis of colorectal cancer patients using three modalities, PET/CT features, histopathology slides, and clinical data, reaching an AUC of 0.83 in the external validation set.

5.4. Reproductive system (n=43)

The reproductive domain encompasses studies focused on the diagnosis and prognosis prediction of malignancies, mainly in the breast (n=32), followed by the prostate (n=5), ovaries, cervix, and placenta (n=5). Unlike other domains, there was no predominant combination of modalities; clinical variables with MRI (n=9) or histopathology (n=6), omics data with histopathology (n=5) or clinical data (n=3) were explored with similar frequency. As in other categories, most studies demonstrated improved performances when integrating multiple modalities (n=33). However, only a few studies (n=9) performed external validation. To train and internally validate their models, some studies employed public datasets, such as the TCGA-BRCA dataset (n=11).

Among interesting studies, Mondol et al. (2024) presented a deep learning framework that fuses image-derived features with genetic and clinical data to perform survival risk stratification of ER+ breast cancer patients. They compared their model to six state-of-the-art models, such as MultiSurv, TransSurv, and MCAT,

demonstrating an improvement in the AUC value ranging from 0.13 to 0.37. Holste et al. (2021) evaluated various methods for fusing MRI and clinical variables in breast cancer classification and achieved an AUC value of 0.989 on an internal validation cohort of 4909 patients, one of the largest validation cohorts we encountered in this domain. Wang et al. (2023b) explored multiple instance learning techniques to combine histopathology with clinical features to predict the prognosis of HER2-positive breast cancer patients and found that simpler modality fusion techniques, such as concatenation, were ineffective in enhancing the model's performance.

5.5. Sensory system (n=25)

In the sensory system, most multimodal models have focused on ophthalmology (n=23), with the remaining studies addressing otology (n=2). There is a distinct focus in ophthalmology on glaucoma (n=7), followed by retinopathy (n=4). Three modalities are the most used in these studies: optical coherence tomography, color fundus photography, and clinical data in different combinations. A high proportion of studies achieved better results employing a multimodal model than unimodal ones (n=19), but only a small number of these papers included external validation (n=3).

We want to highlight Nderitu et al. (2024), who developed a deep learning system to predict one-, two-, and three-year progression of diabetic retinopathy using risk factors and color fundus images, employing more than 160000 eyes for the development of the model and around 28000 and 7000 eyes to internally and externally validate their findings. Zhou et al. (2024b) employed a multimodal approach that combined four imaging modalities with free-text lesion descriptions for uncertainty-aware classification of retinal artery occlusion. Interestingly, their model can also process incomplete data thanks to the Transformer architecture, which is designed to handle input of arbitrary lengths.

5.6. Integumentary system (n=24)

All studies within the integumentary category focused on diagnosing skin lesions, primarily through the fusion of dermatoscopic images and clinical variables (n=12). Most of the studies (n=14) compared multimodal approaches to unimodal baselines and reported improved performance with adding extra modalities, suggesting that factors such as lesion location and patient demographics contribute valuable information. Notably, one paper conducted prospective validation of its findings (Zhu et al., 2024). Only four remaining studies carried out external validation, while all others relied on internal or cross-validation approaches. In contrast to other systems, several publicly available datasets are readily accessible. The most used one is the dataset coming from the ISIC challenge (n=9), followed by HAM10000 (n=4), Seven-point Criteria Evaluation (SPC) (n=3).

Among the most interesting examples of this category is Zhou et al. (2024a), who presented SkinGPT-4, an interactive dermatology diagnostic system based on multimodal large language models and trained on pairs of images and textual descriptions. SkinGPT-4 was evaluated on 150 real-life cases in collaboration with board-certified dermatologists. This system allows users to upload their skin photographs for diagnosis, enabling it to assess the images autonomously, identify the characteristics and classifications of skin conditions, conduct in-depth analyses, and offer interactive treatment recommendations. Other studies include Tang et al. (2022) that proposed two fusion schemes to efficiently integrate dermatoscopic images, clinical photographs, and clinical variables for the diagnosis of eight types of skin lesions, and Zhu et al. (2024) that demonstrated that their multimodal model, which incorporated clinical images and high-frequency ultrasound, performed on par or better than dermatologists in diagnosing seventeen different skin diseases.

5.7. Cardiovascular (n=20)

Research in the cardiovascular domain focused exclusively on diagnosis (n=15), with some studies also concentrating on survival, treatment response, and disease progression prediction (n=4). In all but three studies, the proposed models incorporated clinical

variables with a second modality, often in conjunction with radiology imaging (n=10). Most studies obtained better results when comparing unimodal models with multimodal ones(n=15). However, in this case, only three studies externally validated their results. Some publicly available datasets were employed, including MIMIC, JSRT Database, Montgomery County X-ray Set, and data from the Gene Expression Omnibus (GEO) and UK Biobank.

One interesting example in this category is the study of Jothi Prakash et al. (2024) that introduced an Attention-Based Cross-Modal (ABCM) transfer learning framework to predict cardiovascular disease, merging diverse data types, including clinical records, medical imaging, and genetic information through an attention-driven mechanism. They reached an AUC value of 0.97 on the validation set, significantly surpassing traditional single-source models.

5.8. Urinary system (n=11)

In the urogenital system, the majority of studies focused on the diagnosis of kidney (n=8) disorders, with fewer addressing bladder (n=2) and adrenal gland (n=1) conditions. These multimodal studies were employed for diagnosis (n=7) or survival prediction (n=4) predominantly on oncological malignancies (n=9) with fewer on renal artery stenosis and aiming at performing report generation. Most studies employed CT images combined with a second or third modality (n=9), such as clinical data, histopathology images, and omics. Eight studies demonstrated improved performance with multimodal models compared to unimodal ones. Similarly to other systems, only a small number (2/11) externally validated their results, while most used internal or cross-validation approaches. A few public datasets are available for this category: The Cancer Genome Atlas (TCGA) and The Cancer Imaging Archive (TCIA).

One study focused on renal diseases developed a cross-modal system to create a prognostic model for clear renal cell carcinoma (Ning et al., 2020). This model integrates deep features extracted from computed tomography and histopathological images with eigengenes derived from genomic data, demonstrating a significant ability to stratify patients into high- and low-risk categories for disease progression.

5.9. Musculoskeletal system (n=9)

In the musculoskeletal system, studies focused on diagnosing bone (n=6) diseases, with fewer addressing teeth and muscle conditions, focusing on diagnosing pathological gait, osteoarthritis, deep caries and pulpitis, bone fractures and aging, and soft tissue sarcoma. These studies mainly employed radiology images with clinical data (n=6) or unstructured text data (n=2). Most studies (n=6) demonstrated improved performance with multimodal models compared to unimodal ones. However, no studies externally validated their results, mainly employing internal validation approaches. This category used two public datasets from the Osteoarthritis Initiative (OAI) and the Pediatric Bone Age Challenge.

Li et al. (2023) employed the most extensive validation set of the category, including more than 24000 samples, to diagnose Prosthetic joint infection from CT scan and patients' clinical data using a Unidirectional Selective Attention mechanism and a graph convolutional network and reaching an AUC value of 0.96. Schilcher et al. (2024) included a multicentre dataset of 72 Swedish radiology departments to develop a multimodal model based on X-ray imaging and clinical information for detecting atypical femur fractures.

5.10. Multiple systems (n=27)

This category includes papers that evaluated the multimodal system on multiple organs belonging to different systems. All these studies benefited from available public datasets, confirming that publicly available data is essential for developing and assessing multimodal AI systems. Most of the studies in this category evaluated the performance of their models across diseases affecting up to five organs (n=20). Four studies were distinguished by the extensive range of organs on which they assessed their models. Azher et al. (2023) developed an interpretable multimodal modeling framework that combines DNA methylation, gene

expression, and histopathology for the prognostication of eight types of cancers. Chen et al. (2022) developed a multimodal deep learning model to jointly examine pathology whole-slide images and molecular profile data from 14 cancer types to predict outcomes and discover prognostic features correlating with poor and favorable outcomes. Notably, a vital explainability component was incorporated with heatmaps for histopathology and SHAP values for genomic markers. Cheerla and Gevaert (2019) constructed a multimodal neural network-based model to predict the survival of patients for 20 different cancer types using clinical data, mRNA expression data, microRNA expression data, and histopathology whole slide images (WSIs). Lastly, Vale-Silva and Rohr (2021) presented a multimodal deep learning method for long-term pan-cancer survival prediction that was applied to data from 33 different cancer types, the highest number of organs of the category. These four studies demonstrated the reliability of their models on a wide range of cancer types to do that interestingly, all four benefited from the publicly available data of TCGA.

5.11. Miscellaneous (n=15)

This category includes papers that do not fit into the categories above, such as diagnosis, recurrence, and survival prediction of patients with thyroid carcinoma, prediction of Type II diabetes, severe acute pancreatitis, or immunotherapy response for diffuse large B-cell lymphoma, abnormality detection from chest images, fetal birth weight prediction, and other specific tasks that did not fit the categories above. However, some studies provide novel or unique contributions to multimodal model development. Some interesting examples include Kim and Shin (2023) that proposed a model employing an autoencoder with multiple encoders to extract comprehensive features from hormonal and pathological data and demonstrated that the proposed model significantly improves performance when compared to unimodal models when predicting the recurrence probability of thyroid cancer patients. Lee et al. (2024a) explored unsupervised learning by combining histopathology and clinical data, followed by knowledge distillation to derive a unimodal histopathology model for predicting immunochemotherapy response of patients with Diffuse large B-cell lymphoma. Khader et al. (2023) developed a transformer-based neural network architecture that integrates multimodal patient data, including both imaging and non-imaging, to diagnose up to 25 pathologic conditions and showed that the multimodal model significantly improves diagnostic accuracy when using both chest radiographs and clinical parameters compared to imaging alone and clinical data alone. Finally, Cai et al. (2023) presented a pre-trained multilevel fusion network that combines Vision-conditioned reasoning and Bilinear attentions to enhance feature extraction from medical images and questions. By incorporating Contrastive Language-Image Pre-training (CLIP) and stacked attention layers, their model reduces language bias and improves accuracy. Experiments on three benchmark datasets showed that the proposed model surpasses state-of-the-art models.

6. Towards clinical implementation

In addition to reviewing the landscape of multimodal AI models in research, we investigated to what extent these potential performance boosts can be achieved by clinicians today. To this end, we searched the FDA database of AI/ML-enabled medical devices and the Health AI register for multimodal AI models that obtained either FDA- or CE-clearance. Since these databases contain 950 and 213 entries, we performed an automated preselection of models that may incorporate multiple modalities. Specifically, for each entry in the FDA database, we automatically retrieved the publicly available 510(k) premarket notification summary and searched for any mention of ”multi-modal” or ”multimodal”. This resulted in 38 hits, all manually inspected for the same inclusion criteria used in the literature search. Although one product was found to fit the multimodality criterion by incorporating EEG, neurocognitive test scores, and clinical symptoms to assess the need for CT imaging after potential structural brain injury (Hanley et al., 2017), no deep neural networks were involved, which would place this out of scope for the current review.

A similar strategy was employed for entries in the Health AI register. When available, a published validation study was retrieved and scanned for occurrences of the word ”multimodal” or ”multi-modal.” Although a validation study could be retrieved for 132/213 products, none of these studies returned a hit for our multimodal keywords. Combining our findings from the Food and Drug Administration (FDA) database and AI health register, the promising multimodal AI models showcased in this review have yet to make their way to the clinic.

However, despite the lack of FDA— or CE-certified multimodal AI models, we identified two papers during our literature search that demonstrated potential for transitioning multimodal AI models from research to clinical practice. First, Esteva et al. (2022) describes a multimodal model that integrates digital pathology slides and clinical variables for risk stratification in prostate cancer. Specifically, the model employs a ResNet50 pretrained in a self-supervised manner to extract a feature vector from the pathology slides and subsequently concatenates this image feature vector with a clinical variable vector for the final prediction. Compared to the unimodal imaging model, the multimodal obtained a notably higher AUC of 0.837 vs. 0.779 for predicting distant metastasis after 5 years. In addition, this model was externally validated in a phase 3 trial (NRG/RTOG 9902) and demonstrated significant and independent association with prognostic factors compared to current methods such as the National Comprehensive Cancer Network (NCCN) high-risk features (Ross et al.,
2024). Given this favorable head-to-head performance comparison with current clinical predictive methods, it has since been incorporated into the NCCN guidelines. Although this model is not FDA-certified, clinicians in the United States of America can order it as part of the regular prognostic workup of their patients.

Second, Lee et al. (2022) focused on developing and validating a multimodal prognosis and treatment selection model in COVID-19 patients. This multimodal model employed a previously validated and CE-approved chest X-ray abnormality detection model (Nam et al., 2021) to infer imaging features from chest X-rays and fused this with clinical and laboratory data to predict prognosis and the required interventions for patients with COVID-19. Retrospective validation on 2282 COVID-19 patients from 13 medical centers demonstrated that the multimodal model significantly outperformed the unimodal imaging model with an AUC of 0.854 vs. 0.770, respectively. A feature importance analysis revealed that, among others, clinical variables such as age and dyspnea played an essential role in the multimodal model’s prediction. Although the multimodal model is unavailable for clinicians today, this study proved that current CE-certified unimodal AI models can be incorporated into a multimodal model pipeline to improve performance.

As briefly mentioned in the introduction, medical data is siloed, with each discipline having its own data storage systems. Examples are the Picture Archiving and Communication System (PACS) for radiology, Image Management System (IMS) for pathology, and Electronic Health Records (EHR) for clinical information. Of course, this is an issue for data curation and collection but also a huge issue for implementation, as AI would have to run on data sourced from different systems that do not interoperate.

Additionally, previous studies have demonstrated that by utilizing only a little background information about participants, an adversary could re-identify those in large datasets (Narayanan and Shmatikov, 2008). As the amount of multimodal data collected per patient increases, phenotyping accuracy improves, and significant privacy concerns rise, as the richer data can lead to the re-identification of individuals within large datasets. Managing these privacy issues is crucial to protect patient confidentiality while benefiting from the enhanced insights provided by comprehensive multimodal data.

Lastly, the development of explainable AI (XAI) is of importance to improve trustworthiness by addressing the limitations of AI’s ”black box” nature (Pahud De Mortanges et al., 2024). Explainable AI ensures that end-users can understand and trust the AI’s decision-making process, which is crucial for the widespread adoption and usability of these technologies in clinical practice (Kline

et al., 2022). Prioritizing these strategies will help bridge the gap between research and practical, commercially available multimodal AI solutions in healthcare.

7. Discussion

The exponential growth in multimodal AI models highlights the recent research efforts of multimodal data integration in healthcare. Although multimodal AI development poses unique challenges, the increasing research output in the field will inevitably lead to overcoming some of these challenges. Importantly, in line with a previous review (Kline et al., 2022), our review revealed that integrating multiple data modalities leads to notable performance boosts. In the remainder of this section, we will highlight key takeaways of our analysis and provide recommendations for future research directions.

The general state of the multimodal medical AI development field was summarized in section 3. This overview revealed significant disparities in multimodal AI development across various medical disciplines, tasks, and data domains. Notably, multimodal AI development commonly finds applications in the nervous and respiratory systems. Conversely, applications in the musculoskeletal system or urinary system were scarce. Similar disparities are evident in the realm of data modality combinations. The integration of radiology imaging data and text data greatly outnumbered other combinations and was investigated in almost half of all papers in this review. On the other hand, integrating radiology and pathology data into multimodal frameworks seems to present more significant challenges. Furthermore, substantial variations were observed in the model's application for a given medical task. Most models were aimed at automated diagnosis, whereas prediction of disease progression or survival was less common.

Irrespective of the specific medical discipline or application, these discrepancies consistently point to a common factor: high-quality public datasets' availability (or lack thereof). The increased complexity associated with curating multimodal datasets, often requiring data collection from diverse sources (e.g., electronic health records or physical pathology slides) and departments (e.g., radiology and pathology), currently presents a significant bottleneck in model development. However, collecting, curating, and harmonizing detailed, diverse, and comprehensive annotated datasets representing various demographic factors are essential (Acosta et al., 2022). The richness and quality of such data are paramount for training effective AI models, as they ensure robustness, generalizability, and accuracy in real-world applications. This process, however, faces challenges such as high costs associated with elaborate patient characterization and longitudinal follow-up, which escalate with increasing participant numbers, making automated data collection methods necessary to remain financially sustainable (Acosta et al., 2022).

However, this challenge also presents an opportunity for researchers and clinicians to make impactful contributions to the field, as establishing new multimodal datasets can substantially accelerate AI development in a specific domain. For instance, the publicly available ADNI (Alzheimer's Disease Neuroimaging Initiative) dataset has facilitated the development of more models for Alzheimer's diagnosis (n=45) than all studies combined in the musculoskeletal and urinary system. Apart from enabling domain-specific models, these public datasets can stimulate research tackling general challenges associated with multimodal AI development. When datasets encompass as many as four distinct modalities (i.e. CT imaging, pathology slides, genomic data and text data), as seen in the TCGA-GBM dataset (The Cancer Genome Atlas Glioblastoma Multiforme Collection), this allows extensive investigations on the effects of missing data, model uncertainty, and model explainability with various data modalities (Cui et al., 2022b; Braman et al., 2021; Hao et al., 2020). Hence, we believe that a community effort towards multimodal dataset development can have a strong propelling effect on the progress of the field.

The technical design choices associated with processing and fusing data from diverse modalities were covered in sections 4.3 and 4.4. Our analysis revealed

that most studies prioritize the development of effective modality fusion methods over designing novel encoder architectures for optimal processing of different modalities. This was primarily evidenced by the comparatively large proportion of studies employing pretrained encoders for feature extraction from other modalities, after which the main contribution consisted of a novel, generally intermediate-level, fusion method. We posit that this observed focus on fusion methods rather than data encoders stems from the nature of multimodal datasets, which are typically smaller in scale than their unimodal counterparts. By leveraging encoders pre-trained on large unimodal datasets and occasionally fine-tuning them on the available multimodal data, researchers can achieve robust performance despite the relative scarcity of multimodal data (Wang et al., 2023a; Zhang et al., 2023).

Moreover, the great diversity of modalities in the medical field with their respective intricacies seems to warrant a thorough investigation of optimal data fusion methodologies. A robust fusion method is critical for reliable model predictions, especially in light of the potential absence of specific modalities during inference. The emergence of (medical) foundational models (Moor et al., 2023), which are generally multimodal, is expected to accelerate research in optimal data fusion further. Since these foundational models can further improve the availability of pre-trained solid encoders, this may shift the research focus to designing optimal feature fusion methods. Where we initially saw the main focus on late fusion techniques, where existing unimodal models could straightforwardly be combined with a structured non-imaging modality, newer work moves towards earlier fusion, with most early-fusion papers, which would allow more cross-modality learning, appearing in the past two years. Early fusion has the theoretical advantage that information from the other modalities can improve each modality's encoder, thus creating a 1 + 1 > 2 scenario. However, this is underexplored in the current body of work, and it indicates a promising direction for future research.

Similarly, the development of stronger data encoders may also help tackle the challenge of handling missing data during training and inference. Although the majority of studies in this review discarded entries with missing modalities in their dataset, some innovative approaches leveraged strong encoders to generate learnable embeddings for missing modalities (Chen et al., 2024) or employed encoders with flexible input lengths to handle varying amounts of data (Zhou et al., 2024b).

The clinical applicability of current multimodal AI models, as discussed in subsection 6, reveals that the readiness of these models for clinical implementation lags behind that of unimodal models. We believe this is influenced by the same challenges present in commercializing unimodal AI models, but amplified by the multimodal nature of these models. Although unimodal models require strong, preferably external, validation before deployment, this is even more prudent for multimodal AI models. However, obtaining cross-departmental multimodal data for external validation can be challenging. In addition, since multimodal models consume larger amounts of data during deployment, there may be a greater risk for model drift when one of the data modalities changes. Although extensive validation can investigate the effect of slight modality changes, such as using slightly different staining in a multimodal model incorporating histopathology imaging data, this may lead to a combinatorial explosion when fusing more data modalities. A more optimistic foreseeable future could be that multiple modalities can serve a stabilizing purpose, where a noisy additional modality is unlikely to influence the final prediction adversely. However, we note that fusion of more than two modalities was encountered in only 59/432 (14%) papers, indicating that more evidence is needed to corroborate or disprove any of the aforementioned scenarios. Another amplified challenge in commercializing multimodal AI models is the call for explainable AI. Although this is already challenging for unimodal models, explainability in multimodal AI models likely requires explainability for each component (i.e., Shapley values for tabular data), but perhaps more importantly, some degree of explainability of how information from different sources is combined. Such an explainable multimodal AI model may also alleviate the

aforementioned risk of performance degradation under data drift for one of the data modalities. Given these challenges, we believe that developing public multimodal datasets, preferably incorporating more than two modalities, will be an essential stepping stone for advancing the field.

8. Conclusion

In conclusion, this review provides one of the most comprehensive overviews of multimodal AI development, spanning various medical disciplines, tasks, and data domains. Although substantial evidence exists that multimodal AI models will incur significant performance boosts by taking a broader view of the patient, their development poses novel challenges. We hope this review elucidated some of these challenges, but more importantly, potential solutions to guide the field in the coming years.

CRediT authorship contribution statement

Daan Schouten: Conceptualization, Methodology, Formal Analysis, Data Curation, Visualization, Writing original draft. Giulia Nicoletti: Conceptualization, Methodology, Formal Analysis, Data Curation, Visualization, Writing - original draft. Bas Dille: Conceptualization, Methodology, Formal Analysis, Data Curation, Visualization, Writing - original draft.
Catherine Chia: Conceptualization, Methodology, Formal Analysis, Data Curation, Visualization, Writing
- original draft. Pierpaolo Vendittelli: Methodology, Formal Analysis, Data Curation, Visualization, Writing
- original draft. Megan Schuurmans: Methodology, Formal Analysis, Data Curation, Visualization, Writing
- original draft. Geert Litjens: Conceptualization,
Writing - Review & Editing, Supervision. Nadieh Khalili: Conceptualization, Writing - Review & Editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This research has been funded by the Dutch Research Council, the European Union Horizon Europe Program, and the Hanarth Fund. We acknowledge Mr. H. J. M. Roels's financial support through a donation to Oncode Institute.

Supplementary data

Supplementary material related to this article can be found online.

References

Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J., 2022. Multimodal biomedical AI. Nature Medicine 28, 1773–1784. doi:10.1038/s41591-022-01981-2.

Akramifard, H., Balafar, M.A., Razavi, S.N., Ramli, A.R., 2021. Early detection of alzheimer's disease based on clinical trials, three-dimensional imaging data, and personal information using autoencoders. Journal of Medical Signals & Sensors 11, 120–130. doi:10.4103/jmss.JMSS_11_20.

Azher, Z.L., Suvarna, A., Chen, J.Q., Zhang, Z., Christensen, B.C., Salas, L.A., Vaickus, L.J., Levy, J.J., 2023. Assessment of emerging pretraining strategies in interpretable multimodal deep learning for cancer prognostication. BioData Mining 16, 23. doi:10.1186/s13040-023-00338-w.

Braman, N., Gordon, J.W.H., Goossens, E.T., Willis, C., Stumpe, M.C., Venkataraman, J., 2021. Deep orthogonal fusion: Multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing. pp. 667–677. doi:10.1007/ 978-3-030-87240-3_64.

Cai, L., Fang, H., Li, Z., 2023. Pre-trained multilevel fuse network based on vision-conditioned reasoning and bilinear attentions for medical image visual question answering. The Journal of Supercomputing 79, 13696–13723. doi:10.1007/s11227-023-05195-2.

Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J.,´ Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 9650–9660.

Caruso, C.M., Guarrasi, V., Cordelli, E., Sicilia, R., Gentile, S., Messina, L., Fiore, M., Piccolo, C., Beomonte Zobel, B., Iannello, G., Ramella, S., Soda, P., 2022. A multimodal ensemble driven by multiobjective optimisation to predict overall survival in non-small-cell lung cancer. Journal of Imaging 8, 298. doi:10.3390/jimaging8110298.

Cheerla, A., Gevaert, O., 2019. Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics 35, i446–i454. doi:10.1093/bioinformatics/btz342.

Chen, R.J., Lu, M.Y., Williamson, D.F., Chen, T.Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., Mahmood, F., 2022. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell 40, 865–878.e6. doi:10.1016/j.ccell.2022.07.004.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR. pp. 1597–1607.

Chen, Z., Chen, Y., Sun, Y., Tang, L., Zhang, L., Hu, Y., He, M., Li, Z., Cheng, S., Yuan, J., Wang, Z., Wang, Y., Zhao, J., Gong, J., Zhao, L., Cao, B., Li, G., Zhang, X., Dong, B., Shen, L., 2024. Predicting gastric cancer response to anti-HER2 therapy or anti-HER2 combined immunotherapy based on multi-modal data. Signal Transduction and Targeted Therapy 9, 222. doi:10.1038/s41392-024-01932-y.

Cui, C., Asad, Z., Dean, W.F., Smith, I.T., Madden, C., Bao, S., Landman, B.A., Roland, J.T., Coburn, L.A., Wilson, K.T., Zwerner, J.P., Zhao, S., Wheless, L.E., Huo, Y., 2022a. Multi-modal learning with missing data for cancer diagnosis using histopathological and genomic data. Medical Imaging 2022: Computer-Aided Diagnosis 471, 50. doi:10. 1117/12.2612318.

Cui, C., Liu, H., Liu, Q., Deng, R., Asad, Z., Wang, Y., Zhao, S., Yang, H., Landman, B.A., Huo, Y., 2022b. Survival prediction of brain cancer with incomplete radiology, pathology, genomic, and demographic data, in: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Springer Nature Switzerland. pp. 626–635. doi:10. 1007/978-3-031-16443-9_60.

Cui, H., Zhao, Y., Xiong, S., Feng, Y., Li, P., Lv, Y., Chen, Q., Wang, R., Xie, P., Luo, Z., Cheng, S., Wang, W., Li, X., Xiong, D., Cao, X., Bai, S., Yang, A., Cheng, B., 2024. Diagnosing Solid Lesions in the Pancreas With Multimodal Artificial Intelligence. JAMA Network Open 7, e2422454. doi:10.1001/jamanetworkopen.2024.22454.

Dolci, G., Rahaman, M., Galazzo, I.B., Cruciani, F., Abrol, A., Chen, J., Fu, Z., Duan, K., Menegaz, G., Calhoun, V., 2023. Deep generative transfer learning predicts conversion to alzheimer's disease from neuroimaging genomics data, in: 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing

Workshops (ICASSPW), pp. 1–5. doi:10.1109/ICASSPW59220.2023.10193683.

Esteva, A., Feng, J., van der Wal, D., Huang, S.C., Simko, J.P., DeVries, S., Chen, E., Schaeffer, E.M., Morgan, T.M., Sun, Y., Ghorbani, A., Naik, N., Nathawani, D., Socher, R., Michalski, J.M., Roach, M., Pisansky, T.M., Monson, J.M., Naz, F., Wallace, J., Ferguson, M.J., Bahary, J.P., Zou, J., Lungren, M., Yeung, S., Ross, A.E., Sandler, H.M., Tran, P.T., Spratt, D.E., Pugh, S., Feng, F.Y., Mohamad, O., 2022. Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials. npj Digital Medicine 5, 1–8. doi:10. 1038/s41746-022-00613-w.

Fan, D., Miao, R., Huang, H., Wang, X., Li, S., Huang, Q., Yang, S., Deng, R., 2024. Multimodal ischemic stroke recurrence prediction model based on the capsule neural network and support vector machine. Medicine 103, e39217. doi:10.1097/md. 0000000000039217.

Feher, A., Bednarski, B., Miller, R.J., Shanbhag, A., Lemley, M., Miras, L., Sinusas, A.J., Miller, E.J., Slomka, P.J., 2024. Artificial Intelligence Predicts Hospitalization for Acute Heart Failure Exacerbation in Patients Undergoing Myocardial Perfusion Imaging. Journal of Nuclear Medicine 65, 768–774. doi:10. 2967/jnumed.123.266761.

Flores, J.E., Claborne, D.M., Weller, Z.D., Webb-Robertson, B.M., Waters, K.M., Bramer, L.M., 2023. Missing data in multi-omics integration: Recent advances through artificial intelligence. Frontiers in artificial intelligence 6, 1098308. URL: , doi:10.3389/frai.2023.1098308.

Gao, R., Li, T., Tang, Y., Xu, K., Khan, M., Kammer, M., Antic, S.L., Deppen, S., Huo, Y., Lasko, T.A., Sandler, K.L., Maldonado, F., Landman, B.A., 2022. Reducing uncertainty in cancer risk estimation for patients with indeterminate pulmonary nodules using an integrated deep learning model. Computers in Biology and Medicine 150, 106113. doi:10.1016/j. compbiomed.2022.106113.

Gao, R., Tang, Y., Xu, K., Kammer, M.N., Antic, S.L., Deppen, S., Sandler, K.L., Massion, P.P., Huo, Y., Landman, B.A., 2021. Deep multi-path network integrating incomplete biomarker and chest CT data for evaluating lung cancer risk, in: Medical Imaging 2021: Image Processing, SPIE. pp. 387–393. doi:10.1117/ 12.2580730.

Ghafoori, M., Hamidi, M., Modegh, R.G., Aziz-Ahari, A., Heydari, N., Tavafizadeh, Z., Pournik, O., Emdadi, S., Samimi, S., Mohseni, A., Khaleghi, M., Dashti, H., Rabiee, H.R., 2023. Predicting survival of iranian COVID-19 patients infected by various variants including omicron from CT scan images and clinical data using deep neural networks. Heliyon 9, e21965. doi:10.1016/j.heliyon.2023.e21965.

Hanley, D., Prichep, L.S., Bazarian, J., Huff, J.S., Naunheim, R., Garrett, J., Jones, E.B., Wright, D.W., O'Neill, J., Badjatia, N., Gandhi, D., Curley, K.C., Chiacchierini, R., O'Neil, B., Hack, D.C., 2017. Emergency department triage of traumatic head injury using a brain electrical activity biomarker: A multisite prospective observational validation trial. Academic Emergency Medicine 24, 617–627. doi:10.1111/ acem.13175.

Hao, J., Kosaraju, S.C., Tsaku, N.Z., Song, D.H., Kang, M., 2020. PAGE-net: Interpretable and integrative deep learning for survival analysis using histopathological images and genomic data 25, 355–366.

Holste, G., Partridge, S.C., Rahbar, H., Biswas, D., Lee, C.I., Alessio, A.M., 2021. End-to-end learning of fused image and non-image features for improved breast cancer classification from MRI, in: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 3287–3296. doi:10. 1109/ICCVW54120.2021.00368.

Huang, S.C., Pareek, A., Zamanian, R., Banerjee, I., Lungren, M.P., 2020. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary

embolism detection. Scientific Reports 10, 22147. doi:10.1038/s41598-020-78888-w.

Jothi Prakash, V., Arul Antran Vijay, S., Ganesh Kumar, P., Karthikeyan, N., 2024. A novel attention-based cross-modal transfer learning framework for predicting cardiovascular disease. Computers in Biology and Medicine 170, 107977. doi:10.1016/j.compbiomed.2024.107977.

Jung, H.S., Lee, E.J., Chang, D.I., Cho, H.J., Lee, J., Cha, J.K., Park, M.S., Yu, K.H., Jung, J.M., Ahn, S.H., Kim, D.E., Lee, J.H., Hong, K.S., Sohn, S.I., Park, K.P., Kwon, S.U., Kim, J.S., Chang, J.Y., Kim, B.J., Kang, D.W., 2024. A Multimodal Ensemble Deep Learning Model for Functional Outcome Prognosis of Stroke Patients. Journal of Stroke 26, 312–320. doi:10.5853/jos.2023.03426.

Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G., 2019. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE Journal of Biomedical and Health Informatics 23, 538–546. doi:10.1109/JBHI.2018.2824327.

Kayikci, S., Khoshgoftaar, T.M., 2023. Breast cancer prediction using gated attentive multimodal deep learning. Journal of Big Data 10, 62. doi:10.1186/s40537-023-00749-w.

Keicher, M., Burwinkel, H., Bani-Harouni, D., Paschali, M., Czempiel, T., Burian, E., Makowski, M.R., Braren, R., Navab, N., Wendler, T., 2023. Multimodal graph attention network for COVID-19 outcome prediction. Scientific Reports 13, 19539. doi:10. 1038/s41598-023-46625-8.

Ketabi, S., Agnihotri, P., Zakeri, H., Namdar, K., Khalvati, F., 2023. Multimodal learning for improving performance and explainability of chest x-ray classification, in: Celebi, M.E., Salekin, M.S., Kim, H., Albarqouni, S., Barata, C., Halpern, A., Tschandl, P., Combalia, M., Liu, Y., Zamzmi,
G., Levy, J., Rangwala, H., Reinke, A., Wynn, D., Landman, B., Jeong, W.K., Shen, Y., Deng, Z., Bakas, S., Li, X., Qin, C., Rieke, N., Roth, H., Xu, D. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops, Springer Nature Switzerland. pp. 107–116. doi:10.1007/978-3-031-47401-9_11.

Khader, F., Muller-Franzes, G., Wang, T., Han, T.,¨ Tayebi Arasteh, S., Haarburger, C., Stegmaier, J., Bressem, K., Kuhl, C., Nebelung, S., Kather, J.N., Truhn, D., 2023. Multimodal deep learning for integrating chest radiographs and clinical parameters: A case for transformers. Radiology 309, e230806. doi:10.1148/radiol.230806.

Kim, H., Park, C., Hoon Kim, J., Jang, S., Lee, H.K., 2024. Multimodal Reinforcement Learning for Embedding Networks and Medication Recommendation in Parkinson's Disease. IEEE Access 12, 74251–74267. doi:10.1109/access.2024.3405009.

Kim, J., Shin, H., 2023. Prediction of recurrence probability of thyroid cancer patients using similarity loss based multi-modal autoencoder, in: 2023 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 358–360. doi:10.1109/BigComp57234.2023.00082.

Kline, A., Wang, H., Li, Y., Dennis, S., Hutch, M., Xu, Z., Wang, F., Cheng, F., Luo, Y., 2022. Multimodal machine learning in precision health: A scoping review. npj Digital Medicine 5, 1–14. doi:10.1038/ s41746-022-00712-8.

Kolk, M.Z.H., Ruiperez-Campillo, S., Allaart, C.P.,´
Wilde, A.A.M., Knops, R.E., Narayan, S.M., Tjong, F.V.Y., 2024. Multimodal explainable artificial intelligence identifies patients with non-ischaemic cardiomyopathy at risk of lethal ventricular arrhythmias. Sci Rep 14, 14889. doi:10.1038/s41598-024-65357-x.

Krones, F., Marikkar, U., Parsons, G., Szmul, A., Mahdi, A., 2025. Review of multimodal machine learning

approaches in healthcare. Information Fusion 114, 102690. doi:10.1016/j.inffus.2024.102690.

Kumar, S., Ivanova, O., Melyokhin, A., Tiwari, P., 2023. Deep-learning-enabled multimodal data fusion for lung disease classification. Informatics in Medicine Unlocked 42, 101367. doi:10.1016/j.imu.2023. 101367.

Lee, G., Nho, K., Kang, B., Sohn, K.A., Kim, D., 2019. Predicting alzheimer's disease progression using multi-modal deep learning approach. Scientific Reports 9, 1952. doi:10.1038/s41598-018-37769-z.

Lee, J.H., Ahn, J.S., Chung, M.J., Jeong, Y.J., Kim, J.H., Lim, J.K., Kim, J.Y., Kim, Y.J., Lee, J.E., Kim, E.Y., 2022. Development and validation of a multimodal-based prognosis and intervention prediction model for COVID-19 patients in a multicenter cohort. Sensors 22, 5007. doi:10. 3390/s22135007.

Lee, J.H., Song, G., Lee, J., Kang, S., Moon, K.M., Choi, Y., Shen, J., Noh, M., Yang, D., 2024a. Prediction of immunochemotherapy response for diffuse large <scp>b</scp>-cell lymphoma using artificial intelligence digital pathology. The Journal of Pathology: Clinical Research 10, e12370. doi:10. 1002/2056-4538.12370.

Lee, S., Cho, Y., Ji, Y., Jeon, M., Kim, A., Ham, B.J., Joo, Y.Y., 2024b. Multimodal integration of neuroimaging and genetic data for the diagnosis of mood disorders based on computer vision models. Journal of Psychiatric Research 172, 144–155. doi:10. 1016/j.jpsychires.2024.02.036.

Lei, B., Li, Y., Fu, W., Yang, P., Chen, S., Wang, T., Xiao, X., Niu, T., Fu, Y., Wang, S., Han, H., Qin, J., 2024. Alzheimer's disease diagnosis from multi-modal data via feature inductive learning and dual multilevel graph neural network. Medical Image Analysis 97, 103213. doi:10.1016/j.media.2024.103213.

Li, R., Yang, F., Liu, X., Shi, H., 2023. HGT: A hierarchical GCN-based transformer for multimodal periprosthetic joint infection diagnosis using computed tomography images and text. Sensors 23, 5795. doi:10.3390/s23135795.

Li, Z., Ogino, M., 2020. Augmented radiology: Patient-wise feature transfer model for glioma grading, in: Albarqouni, S., Bakas, S., Kamnitsas, K., Cardoso, M.J., Landman, B., Li, W., Milletari, F., Rieke, N., Roth, H., Xu, D., Xu, Z. (Eds.), Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning, Springer International Publishing. pp. 23–30. doi:10.1007/ 978-3-030-60548-3_3.

Lim, W.S., Ho, H.Y., Ho, H.C., Chen, Y.W., Lee, C.K., Chen, P.J., Lai, F., Jang, J.S.R., Ko, M.L., 2022. Use of multimodal dataset in AI for detecting glaucoma based on fundus photographs assessed with OCT: focus group study on high prevalence of myopia. BMC Medical Imaging 22, 206. doi:10.1186/ s12880-022-00933-z.

Lipkova, J., Chen, R.J., Chen, B., Lu, M.Y., Barbieri, M., Shao, D., Vaidya, A.J., Chen, C., Zhuang, L., Williamson, D.F.K., Shaban, M., Chen, T.Y., Mahmood, F., 2022. Artificial intelligence for multimodal data integration in oncology. Cancer Cell 40, 1095–1110. doi:10.1016/j.ccell.2022.09. 012.

Liu, F., Yuan, S., Li, W., Xu, Q., Wu, X., Han, K., Wang, J., Miao, S., 2024. Multi-task joint learning network based on adaptive patch pruning for Alzheimer's disease diagnosis and clinical score prediction. Biomedical Signal Processing and Control 95, 106398. doi:10.1016/j.bspc.2024.106398.

Liu, H., Zhao, Y., Yang, F., Lou, X., Wu, F., Li, H., Xing, X., Peng, T., Menze, B., Huang, J., Zhang, S., Han, A., Yao, J., Fan, X., 2022. Preoperative prediction of lymph node metastasis in colorectal cancer with deep learning. BME Frontiers 2022, 9860179. doi:10. 34133/2022/9860179.

Liu, L., Liu, S., Zhang, L., To, X.V., Nasrallah, F., Chandra, S.S., 2023. Cascaded multi-modal mixing transformers for alzheimer's disease classification with incomplete data. NeuroImage 277, 120267. doi:10.1016/j.neuroimage.2023.120267.

Lobbestael, G., 2023. Dedupendnote. URL: .

Lopez, K., Fodeh, S.J., Allam, A., Brandt, C.A., Krauthammer, M., 2020. Reducing annotation burden through multimodal learning. Frontiers in Big Data 3. doi:10.3389/fdata.2020.00019.

Ma, X., Jia, F., 2020. Brain tumor classification with multimodal MR and pathology images, in: Crimi, A., Bakas, S. (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer International Publishing. pp. 343–352. doi:10.1007/978-3-030-46643-5_34.

Machado Reyes, D., Chao, H., Hahn, J., Shen, L., Yan, P., Alzheimers Dis, N., 2024. Identifying Progression-Specific Alzheimer's Subtypes Using Multimodal Transformer. JOURNAL OF PERSONALIZED MEDICINE 14, 421. doi:10.3390/jpm14040421.

Mano, M.S., c¸itaku, F.T., Barach, P., 2022. Implementing Multidisciplinary Tumor Boards in Oncology: a Narrative Review. Future Oncology 18, 375–384. doi:10.2217/fon-2021-0471.

Menegotto, A.B., Becker, C.D.L., Cazella, S.C., 2021. Computer-aided diagnosis of hepatocellular carcinoma fusing imaging and structured health data. Health Information Science and Systems 9, 20. doi:10.1007/ s13755-021-00151-x.

Mondol, R.K., Millar, E.K.A., Sowmya, A., Meijering, E., 2024. BioFusionNet: Deep Learning-Based Survival Risk Stratification in ER+ Breast Cancer Through Multifeature and Multimodal Data Fusion. IEEE Journal of Biomedical and Health Informatics 28, 5290–5302. doi:10.1109/jbhi.2024.3418341.

Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P., 2023. Foundation models for generalist medical artificial intelligence. Nature 616, 259–265. doi:10.1038/ s41586-023-05881-4.

Mustafa, E., Jadoon, E.K., Khaliq-uz Zaman, S., Humayun, M.A., Maray, M., 2023. An ensembled framework for human breast cancer survivability prediction using deep learning. Diagnostics 13, 1688. doi:10.3390/diagnostics13101688.

Nam, J.G., Kim, M., Park, J., Hwang, E.J., Lee, J.H., Hong, J.H., Goo, J.M., Park, C.M., 2021. Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs. European Respiratory Journal 57, 2003061. doi:10.1183/ 13993003.03061-2020.

Narayanan, A., Shmatikov, V., 2008. Robust de-anonymization of large sparse datasets, in: 2008 IEEE Symposium on Security and Privacy (sp 2008), pp. 111–125. doi:10.1109/SP.2008.33.

Nderitu, P., Nunez do Rio, J.M., Webster, L., Mann, S., Cardoso, M.J., Modat, M., Hopkins, D., Bergeles, C., Jackson, T.L., 2024. Predicting 1, 2 and 3 year emergent referable diabetic retinopathy and maculopathy using deep learning. Communications Medicine 4, 167. doi:10.1038/s43856-024-00590-z.

Nguyen, H.H., Blaschko, M.B., Saarakkala, S., Tiulpin, A., 2024. Clinically-inspired multi-agent transformers for disease trajectory forecasting from multimodal data. IEEE Transactions on Medical Imaging 43, 529–541. doi:10.1109/TMI.2023.3312524.

Niazi, M.K.K., Parwani, A.V., Gurcan, M.N., 2019. Digital pathology and artificial intelligence. The Lancet Oncology 20, e253–e261. doi:10.1016/ S1470-2045(19)30154-8.

Ning, Z., Pan, W., Chen, Y., Xiao, Q., Zhang, X., Luo, J., Wang, J., Zhang, Y., 2020. Integrative analysis of cross-modal features for the prognosis prediction of clear cell renal cell carcinoma. Bioinformatics 36, 2888–2895. doi:10.1093/bioinformatics/ btaa056.

Ostertag, C., Visani, M., Urruty, T., Beurton-Aimar, M., 2023. Long-term cognitive decline prediction based on multi-modal data using multimodal3dsiamesenet: transfer learning from alzheimer's disease to parkinson's disease. International Journal of Computer Assisted Radiology and Surgery 18, 809–818. doi:10.1007/s11548-023-02866-6.

Ouzzani, M., Hammady, H., Fedorowicz, Z., Elmagarmid, A., 2016. Rayyan-a web and mobile app for systematic reviews. Systematic Reviews 5, 210. doi:10.1186/s13643-016-0384-4.

Pahud De Mortanges, A., Luo, H., Shu, S.Z., Kamath, A., Suter, Y., Shelan, M., Pollinger, A., Reyes, M.,¨ 2024. Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. npj Digital Medicine 7, 195. doi:10.1038/ s41746-024-01190-w.

Palmal, S., Arya, N., Saha, S., Tripathy, S., 2024. Integrative prognostic modeling for breast cancer: Unveiling optimal multimodal combinations using graph convolutional networks and calibrated random forest. Applied Soft Computing 154, 111379. doi:10. 1016/j.asoc.2024.111379.

Parvin, S., Nimmy, S.F., Kamal, M.S., 2024. Convolutional neural network based data interpretable framework for alzheimer's treatment planning. Visual Computing for Industry, Biomedicine, and Art 7, 3. doi:10.1186/s42492-024-00154-x.

Pelka, O., Friedrich, C.M., Nensa, F., Monningho¨ ff, C., Bloch, L., Jockel,¨ K.H., Schramm, S., Sanchez Hoffmann, S., Winkler, A., Weimar, C., Jokisch, M., Alzheimer's Disease Neuroimaging Initiative, 2020. Sociodemographic data and APOE-ε4 augmentation for MRI-based detection of amnestic mild cognitive impairment using deep learning systems. PLOS ONE 15, e0236868. doi:10.1371/journal.pone.0236868.

Qiu, S., Miller, M.I., Joshi, P.S., Lee, J.C., Xue, C., Ni, Y., Wang, Y., De Anda-Duran, I., Hwang, P.H., Cramer, J.A., Dwyer, B.C., Hao, H., Kaku, M.C., Kedar, S., Lee, P.H., Mian, A.Z., Murman, D.L., O'Shea, S., Paul, A.B., Saint-Hilaire, M.H., Alton Sartor, E., Saxena, A.R., Shih, L.C., Small, J.E., Smith, M.J., Swaminathan, A., Takahashi, C.E., Taraschenko, O., You, H., Yuan, J., Zhou, Y., Zhu, S., Alosco, M.L., Mez, J., Stein, T.D., Poston, K.L., Au, R., Kolachalama, V.B., 2022. Multimodal deep learning for alzheimer's disease dementia assessment. Nature Communications 13, 3404. doi:10.1038/ s41467-022-31037-5.

Qiu, Y., Lu, H., Mei, J., Bao, S., Xu, J., 2024. Towards semi-supervised multi-modal rectal cancer segmentation: A large-scale dataset and a multi-teacher uncertainty-aware network. Expert Systems with Applications 255, 124734. doi:10.1016/j.eswa. 2024.124734.

Rahman, T., Chowdhury, M.E.H., Khandakar, A., Mahbub, Z.B., Hossain, M.S.A., Alhatou, A., Abdalla, E., Muthiyal, S., Islam, K.F., Kashem, S.B.A., Khan, M.S., Zughaier, S.M., Hossain, M., 2023. BIO-CXRNET: a robust multimodal stacking machine learning technique for mortality risk prediction of COVID-19 patients using chest x-ray images and clinical data. Neural Computing and Applications 35, 17461–17483. doi:10.1007/s00521-023-08606-w.

Ross, A.E., Zhang, J., Huang, H.C., Yamashita, R., Keim-Malpass, J., Simko, J.P., DeVries, S., Morgan, T.M., Souhami, L., Dobelbower, M.C., McGinnis, L.S., Jones, C.U., Dess, R.T., Zeitzer, K.L., Choi, K., Hartford, A.C., Michalski, J.M., Raben, A., Gomella, L.G., Sartor, A.O., Rosenthal, S.A., Sandler, H.M., Spratt, D.E., Pugh, S.L., Mohamad, O., Esteva, A., Chen, E., Schaeffer, E.M., Tran, P.T., Feng, F.Y., 2024. External validation of a digital pathology-based multimodal artificial intelligence architecture in the NRG/RTOG 9902 phase 3 trial. European Urology Oncology , S2588–9311(24)00029–4doi:10.1016/j. euo.2024.01.004.

Saad, M., He, S., Thorstad, W., Gay, H., Barnett, D., Zhao, Y., Ruan, S., Wang, X., Li, H., 2022. Learning-based

cancer treatment outcome prognosis using multimodal biomarkers. IEEE Transactions on Radiation and Plasma Medical Sciences 6, 231–244. doi:10.1109/trpms.2021.3104297.

Salvi, M., Loh, H.W., Seoni, S., Barua, P.D., Garc´ıa, S., Molinari, F., Acharya, U.R., 2024. Multi-modality approaches for medical support systems: A systematic review of the last decade. Information Fusion 103, 102134. doi:10.1016/j.inffus.2023.102134.

Schilcher, J., Nilsson, A., Andlid, O., Eklund, A., 2024. Fusion of electronic health records and radiographic images for a multimodal deep learning prediction model of atypical femur fractures. Computers in Biology and Medicine 168, 107704. doi:10.1016/j.compbiomed.2023.107704.

Sempionatto, J.R., Lasalde-Ram´ırez, J.A., Mahato, K., Wang, J., Gao, W., 2022. Wearable chemical sensors for biomarker discovery in the omics era. Nature Reviews Chemistry 6, 899–915. doi:10.1038/ s41570-022-00439-w.

Shi, M., Li, X., Li, M., Si, Y., 2023. Attention-based generative adversarial networks improve prognostic outcome prediction of cancer from multimodal data. Briefings in Bioinformatics 24, bbad329. doi:10.1093/bib/bbad329.

Shilo, S., Rossman, H., Segal, E., 2020. Axes of a revolution: challenges and promises of big data in healthcare. Nature Medicine 26, 29–38. doi:10.1038/ s41591-019-0727-5.

Steyaert, S., Pizurica, M., Nagaraj, D., Khandelwal, P., Hernandez-Boussard, T., Gentles, A.J., Gevaert, O., 2023. Multimodal data fusion for cancer biomarker discovery with deep learning. Nature Machine Intelligence 5, 351–362. doi:10.1038/ s42256-023-00633-5.

Taleb, A., Kirchler, M., Monti, R., Lippert, C., 2022. ContIG: Self-supervised multimodal contrastive learning for medical imaging with genetics, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20876–20889. doi:10.1109/CVPR52688.2022.02024.

Tang, P., Yan, X., Nan, Y., Xiang, S., Krammer, S., Lasser, T., 2022. FusionM4net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification. Medical Image Analysis 76, 102307. doi:10.1016/j.media.2021.102307.

Tortora, M., Cordelli, E., Sicilia, R., Nibid, L., Ippolito, E., Perrone, G., Ramella, S., Soda, P., 2023. RadioPathomics: Multimodal learning in non-small cell lung cancer for adaptive radiotherapy. IEEE Access 11, 47563–47578. doi:10.1109/ACCESS.2023.3275126.

Vaidya, P., Bera, K., Gupta, A., Wang, X., Corredor, G., Fu, P., Beig, N., Prasanna, P., Patil, P.D., Velu, P.D., Rajiah, P., Gilkeson, R., Feldman, M.D., Choi, H., Velcheti, V., Madabhushi, A., 2020. CT derived radiomic score for predicting the added benefit of adjuvant chemotherapy following surgery in stage i, II resectable non-small cell lung cancer: a retrospective multicohort study for outcome prediction 2, e116–e128. doi:10.1016/ S2589-7500(20)30002-9.

Vale-Silva, L.A., Rohr, K., 2021. Long-term cancer survival prediction using multimodal deep learning. Scientific Reports 11, 13505. doi:10.1038/ s41598-021-92799-4.

Wang, R., Chen, L.C., Moukheiber, L., Seastedt, K.P., Moukheiber, M., Moukheiber, D., Zaiman, Z., Moukheiber, S., Litchman, T., Trivedi, H., Steinberg, R., Gichoya, J.W., Kuo, P.C., Celi, L.A., 2023a. Enabling chronic obstructive pulmonary disease diagnosis through chest x-rays: A multi-site and multi-modality study. International Journal of Medical Informatics 178, 105211. doi:10.1016/j.ijmedinf.2023.105211.

Wang, Y., Luo, Y., Li, B., Shen, X., 2024a. Multi-modality Fusion Based Lung Cancer Survival Analysis with Self-supervised Whole Slide Image Representation Learning, in: Liu, Q., Wang, H., Ma, Z.,

Zheng, W., Zha, H., Chen, X., Wang, L., Ji, R. (Eds.), Pattern Recognition and Computer Vision, Springer Nature, Singapore. pp. 333–345. doi:10.1007/978-981-99-8558-6_28.

Wang, Y., Zhang, L., Li, Y., Wu, F., Cao, S., Ye, F., 2023b. Predicting the prognosis of HER2-positive breast cancer patients by fusing pathological whole slide images and clinical features using multiple instance learning. Mathematical Biosciences and Engineering 20, 11196–11211. doi:10.3934/mbe.2023496.

Wang, Z., Lin, R., Li, Y., Zeng, J., Chen, Y., Ouyang, W., Li, H., Jia, X., Lai, Z., Yu, Y., Yao, H., Su, W., 2024b. Deep learning-based multi-modal data integration enhancing breast cancer disease-free survival prediction. Precision Clinical Medicine 7, pbae012. doi:10.1093/pcmedi/pbae012.

Wang, Z., Luo, S., Chen, J., Jiao, Y., Cui, C., Shi, S., Yang, Y., Zhao, J., Jiang, Y., Zhang, Y., Xu, F., Xu, J., Lin, Q., Dong, F., 2024c. Multi-modality deep learning model reaches high prediction accuracy in the diagnosis of ovarian cancer. iScience 27, 109403. doi:10.1016/j.isci.2024.109403.

Wu, X., Shi, Y., Wang, M., Li, A., 2023. CAMR: cross-aligned multimodal representation learning for cancer survival prediction. Bioinformatics 39, btad025. doi:10.1093/bioinformatics/btad025.

Xu, L., Wu, H., He, C., Wang, J., Zhang, C., Nie, F., Chen, L., 2022a. Multi-modal sequence learning for alzheimer's disease progression prediction with incomplete variable-length longitudinal data. Medical Image Analysis 82, 102643. doi:10.1016/j.media.2022.102643.

Xu, Y., Liu, X., Pan, L., Mao, X., Liang, H., Wang, G., Chen, T., 2022b. Explainable dynamic multimodal variational autoencoder for the prediction of patients with suspected central precocious puberty 26.

Xue, C., Kowshik, S.S., Lteif, D., Puducheri, S., Jasodanand, V.H., Zhou, O.T., Walia, A.S., Guney, O.B., Zhang, J.D., Pham, S.T., Kaliaev, A., Andreu-Arasa, V.C., Dwyer, B.C., Farris, C.W., Hao, H., Kedar, S., Mian, A.Z., Murman, D.L., O'Shea, S.A., Paul, A.B., Rohatgi, S., Saint-Hilaire, M.H., Sartor, E.A., Setty, B.N., Small, J.E., Swaminathan, A., Taraschenko, O., Yuan, J., Zhou, Y., Zhu, S., Karjadi, C., Alvin Ang, T.F., Bargal, S.A., Plummer, B.A., Poston, K.L., Ahangaran, M., Au, R., Kolachalama, V.B., 2024. AI-based differential diagnosis of dementia etiologies on multimodal data. Nature Medicine , nandoi:10.1038/s41591-024-03118-z.

Yang, J., Ju, J., Guo, L., Ji, B., Shi, S., Yang, Z., Gao, S., Yuan, X., Tian, G., Liang, Y., Yuan, P., 2022. Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. Computational and Structural Biotechnology Journal 20, 333–342. doi:10.1016/j.csbj.2021.12.028.

Yang, L., Wang, X., Guo, Q., Gladstein, S., Wooten, D., Li, T., Robieson, W.Z., Sun, Y., Huang, X., for the Alzheimer's Disease Neuroimaging Initiative, 2021. Deep Learning Based Multimodal Progression Modeling for Alzheimer's Disease. Statistics in Biopharmaceutical Research 13, 337–343. doi:10.1080/19466315.2021.1884129.

Yin, M., Lin, J., Wang, Y., Liu, Y., Zhang, R., Duan, W., Zhou, Z., Zhu, S., Gao, J., Liu, L., Liu, X., Gu, C., Huang, Z., Xu, X., Xu, C., Zhu, J., 2024. Development and validation of a multimodal model in predicting severe acute pancreatitis based on radiomics and deep learning. International Journal of Medical Informatics 184, 105341. doi:10.1016/j.ijmedinf.2024.105341.

Ying, Q., Xing, X., Liu, L., Lin, A.L., Jacobs, N., Liang, G., 2021. Multi-modal data analysis for alzheimer's disease diagnosis: An ensemble model using imagery and genetic features. 2021 43rd Annual International Conference of the IEEE Engineering in Medicine &

Biology Society (EMBC) 542, 3586–3591. doi:10.1109/EMBC46164.2021.9630174.

Yu, Q., Ma, Q., Da, L., Li, J., Wang, M., Xu, A., Li, Z., Li, W., 2024. A transformer-based unified multimodal framework for Alzheimer's disease assessment. Computers in Biology and Medicine 180, 108979. doi:10.1016/j.compbiomed.2024.108979.

Zambrano Chaves, J.M., Wentland, A.L., Desai, A.D., Banerjee, I., Kaur, G., Correa, R., Boutin, R.D., Maron, D.J., Rodriguez, F., Sandhu, A.T., Rubin, D., Chaudhari, A.S., Patel, B.N., 2023. Opportunistic assessment of ischemic heart disease risk using abdominopelvic computed tomography and medical record data: a multimodal explainable artificial intelligence approach. Scientific Reports 13, 21034. doi:10.1038/s41598-023-47895-y.

Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y., 2023. Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Communications 14, 4542. doi:10.1038/s41467-023-40260-7.

Zhi, Z., Elbadawi, M., Daneshmend, A., Orlu, M., Basit, A., Demosthenous, A., Rodrigues, M., 2022. Multimodal diagnosis for pulmonary embolism from EHR data and CT images. 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) 10, 2053–2057. doi:10.1109/EMBC48229.2022.9871041.

Zhou, J., He, X., Sun, L., Xu, J., Chen, X., Chu, Y., Zhou, L., Liao, X., Zhang, B., Afvari, S., Gao, X., 2024a. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. Nature Communications 15, 5649. doi:10.1038/s41467-024-50043-3.

Zhou, Q., Chen, T., Zou, H., Xiao, X., 2024b. Uncertainty-aware incomplete multimodal fusion for few-shot central retinal artery occlusion classification. Information Fusion 104, 102200. doi:10.1016/j.inffus.2023.102200.

Zhou, S., Sun, D., Mao, W., Liu, Y., Cen, W., Ye, L., Liang, F., Xu, J., Shi, H., Ji, Y., Wang, L., Chang, W., 2023. Deep radiomics-based fusion model for prediction of bevacizumab treatment response and outcome in patients with colorectal cancer liver metastases: a multicentre cohort study. eClinicalMedicine 65, 102271. doi:10.1016/j. eclinm.2023.102271.

Zhu, A.Q., Wang, Q., Shi, Y.L., Ren, W.W., Cao, X., Ren, T.T., Wang, J., Zhang, Y.Q., Sun, Y.K., Chen, X.W., Lai, Y.X., Ni, N., Chen, Y.C., Hu, J.L., Mou, L.C., Zhao, Y.J., Liu, Y.Q., Sun, L.P., Zhu, X.X., Xu, H.X., Guo, L.H., 2024. A deep learning fusion network trained with clinical and high-frequency ultrasound images in the multi-classification of skin diseases in comparison with dermatologists: a prospective and multicenter study. eClinicalMedicine 67, 102391. doi:10.1016/j.eclinm.2023.102391.