Multimodal Machine Learning in Image-Based and Clinical Biomedicine: Survey and Prospects

Elisa Warner[1] · Joonsang Lee[1] · William Hsu[2] · Tanveer Syeda-Mahmood[3] · Charles E. Kahn Jr.[4] · Olivier Gevaert[5] · Arvind Rao[1]

Abstract

Machine learning (ML) applications in medical artificial intelligence (AI) systems have shifted from traditional and statistical methods to increasing application of deep learning models. This survey navigates the current landscape of multimodal ML, focusing on its profound impact on medical image analysis and clinical decision support systems. Emphasizing challenges and innovations in addressing multimodal representation, fusion, translation, alignment, and co-learning, the paper explores the transformative potential of multimodal models for clinical predictions. It also highlights the need for principled assessments and practical implementation of such models, bringing attention to the dynamics between decision support systems and healthcare providers and personnel. Despite advancements, challenges such as data biases and the scarcity of "big data" in many biomedical domains persist. We conclude with a discussion on principled innovation and collaborative efforts to further the mission of seamless integration of multimodal ML models into biomedical practice.

Communicated by Paolo Rota.

Dr. Rao as primary corresponding author and Elisa Warner as secondary corresponding author.

✉ Elisa Warner
elisawa@umich.edu

✉ Arvind Rao
ukarvind@med.umich.edu

Joonsang Lee
leejoons@umich.edu

William Hsu
whsu@mednet.ucla.edu

Tanveer Syeda-Mahmood
stf@us.ibm.com

Charles E. Kahn Jr.
ckahn@upenn.edu

Olivier Gevaert
ogevaert@stanford.edu

1   Department of Computational Medicine and Bioinformatics, University of Michigan Ann Arbor, 100 Washtenaw Ave., Ann Arbor, MI 48109, USA

1 Introduction

Machine learning (ML), the process of leveraging algorithms and optimization to infer strategies for solving learning tasks, has enabled some of the greatest developments in artificial intelligence (AI) in the last decade, enabling the automated segmentation or class identification of images, the ability to answer nearly any text-based question, and the ability to generate images never seen before. In biomedical research, many of these ML models are quickly being applied to medical images and decision support systems in conjunction with a significant shift from traditional and statistical methods to increasing application of deep learning models. At the same

2   Department of Medical and Imaging Informatics, University of California Los Angeles, 924 Westwood Blvd Ste 420, Los Angeles, CA 90024, USA

3   Almaden Research Center, IBM, 650 Harry Rd., San Jose, CA 95120, USA

4   Department of Radiology, University of Pennsylvania, 3400 Spruce St., Philadelphia, PA 19104, USA

5   Center for Biomedical Informatics Research, Stanford, 1265 Welch Road, Stanford, CA 94305, USA

time, the importance of both plentiful and well-curated data has become better understood, coinciding as of the time of writing this article with the incredible premise of OpenAI's ChatGPT and GPT-4 engines as well as other generative AI models which are trained on a vast, well-curated, and diverse array of content from across the internet (OpenAI, 2023).

As more data has become available, a wider selection of datasets containing more than one modality has also enabled growth in the multimodal research sphere. Multimodal data is intrinsic to biomedical research and clinical care. While data belonging to a single modality can be conceptualized as a way in which something is perceived or captured in the world into an abstract digitized representation such as a waveform or image, multimodal data aggregates multiple modalities and thus consists of several intrinsically different representation spaces (and potentially even different data geometries). Computed tomography (CT) and positron emission tomography (PET) are specific examples of single imaging modalities, while magnetic resonance imaging (MRI) is an example itself of multimodal data, as its component sequences T1-weighted, T2-weighted, and fluid-attenuated inversion recovery (FLAIR) can each be considered their own unique modalities, since each of the MR sequences measure some different biophysical or biological property. Laboratory blood tests, patient demographics, electrocardiogram (ECG) and genetic expression values are also common modalities in clinical decision models. This work discusses unique ways that differences between modalities have been addressed and mitigated to improve accuracy of AI models in similar ways to which a human would naturally be able to re-calibrate to these differences.

There is conceptual value to building multimodal models. Outside of the biomedical sphere, many have already witnessed the sheer power of multimodal AI in text-to-image generators such as DALL·E 2, DALL·E 3 or Midjourney (Ramesh et al., 2022; Betker et al., 2023; Oppenlaender, 2022), some of whose artful creations have won competitions competing against humans (Metz, 2022). In the biomedical sphere, multimodal models provide potentially more robust and generalizable AI predictions as well as a more holistic approach to diagnosis or prognosis of patients, akin to a more human-like approach to medicine. While a plethora of biomedical AI publications based on unimodal data exist, fewer multimodal models exist due to cost and availability constraints of obtaining multimodal data. However, since patient imaging and lab measurements are decreasing in cost and increasing in availability, the case for building multimodal biomedical AI is becoming increasingly compelling.

With the emergence of readily-available multimodal data comes new challenges and responsibilities for those who use them. The survey and taxonomy from Baltrusaitis et al. (2019) presents an organized description of these new challenges, which can be summarized in Fig. 1: (1) rep-

resentation, (2) fusion, (3) alignment, (4) translation, (5) co-learning. *Representation* often condenses a single modality such as audio or an image to a machine-readable data structure such as a vector, matrix, tensor object, or other geometric form, and is concerned with ways to combine more than one modality into the same representation space. Good multimodal representations are constructed in ways in which relationships and context can be preserved between modalities. Multimodal *fusion* relates to the challenge of how to properly combine multimodal data into a predictive model. In multimodal *alignment*, models attempt to automatically align one modality to another. In a simple case, models could be constructed to align PPG signals taken at a 60Hz sampling frequency with a 240Hz ECG signal. In a more challenging case, video of colonoscopy could be aligned to an image representing the camera's location in the colon. Multimodal *translation* consists of mapping one modality to another. For example, several popular natural language processing (NLP) models attempt to map an image to a description of the image, switching from the imaging domain to a text domain. In translational medicine, image-to-image translation tends to be the most common method, whereby one easily-obtained imaging domain such as CT is converted to a harder-to-obtain domain such as T1-weighted MRI. Lastly, multimodal *co-learning* involves the practice of transferring knowledge learned from one modality to a model or data from a different modality.

In this paper, we use the taxonomical framework from Baltrusaitis et al. (2019) to survey current methods which address each of the five challenges of multimodal learning with a novel focus on addressing these challenges in medical image-based clinical decision support. The aim of this work is to introduce both current and new approaches for addressing each multimodal challenge. We conclude with a discussion on the future of AI in biomedicine and what steps we anticipate could further progress in the field.

2 Multimodal Learning in Medical Applications

In the following section, we reintroduce the five common challenges in multimodal ML addressed in Sect. 1 and discuss modern approaches to each challenge as applied to image-based biomedicine. The taxonomical subcategories of Representation and Fusion are summarized in Fig. 2, while those for Translation, Alignment and Co-learning are summarized in Fig. 3. A table of relevant works by the challenge addressed and data types used are given in Table 1.

2.1 Representation

Representation in machine learning typically entails the challenge of transferring contextual knowledge of a complex

Fig. 1 Challenges in multimodal learning: (1) representation, which concerns how multiple modalities will be geometrically represented and how to represent intrinsic relationships between them; (2) fusion, the challenge of combining multiple modalities into a predictive model; (3) translation, involving the mapping of one modality to another; (4) alignment, which attempts to align two separate modalities spatially or temporally; and (5) co-learning, which involves using one modality to assist the learning of another modality
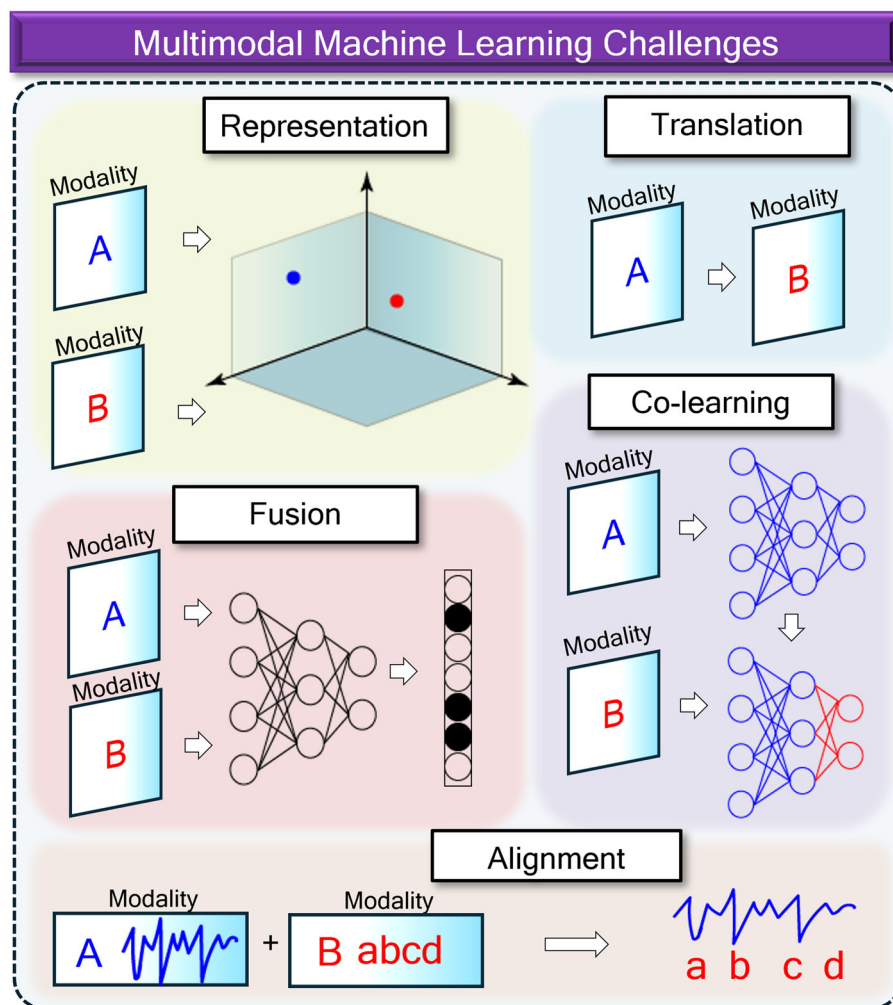
entity such as an image or sound to a mathematically-interpretable or machine-readable format such as a vector or a matrix. Prior to the rise of deep learning, features were engineered in images using techniques such as the aforementioned Scale-Invariant Feature Transform (SIFT) or through methods such as edge detection. Features in audio or other waveform signals such as ECG could be extracted utilizing wavelets or Fourier transform to isolate latent properties of signals and then quantitative values could be derived from morphological patterns in the extracted signal. Multimodal representation challenges venture a step further, consisting of the ability to translate similarities and differences from one modality's representation to another modality's representation. For example, when representing both medical text and CT images, if the vector representations for "skull" and "brain" in medical *text* are closer than those for "skull" and "pancreas", then in a good CT representation, such relationships between vector representations of these structures in the *image* should remain preserved. The derivation of "good" representations in multimodal settings have been outlined in

Bengio et al. (2013) and extended by Srivastava and Salakhutdinov (2014).

It is crucial to acknowledge that representation becomes notably challenging when dealing with more abstract concepts. In a unimodal context, consider the task of crafting representations from an image. Beyond pixel intensities, these representations must encapsulate contextual and semantically-proximate information from the image. A simplistic model may fail to encode context adequately, discerning insufficient distinctions between a foreground and background to represent nuanced visual-semantic concepts. Achieving such subtleties in representations, particularly in abstract contexts, poses increased challenges compared to quantifying similarities and differences in less-nuanced data such as cell counts or gene expression.

Prior to delving into multimodal representations, it is instructive to elucidate strategies for crafting proficient unimodal representations, as multimodal approaches often involve combining or adapting multiple unimodal methods. For images, *pretrained networks* are a common approach for transforming images into good vector representations.
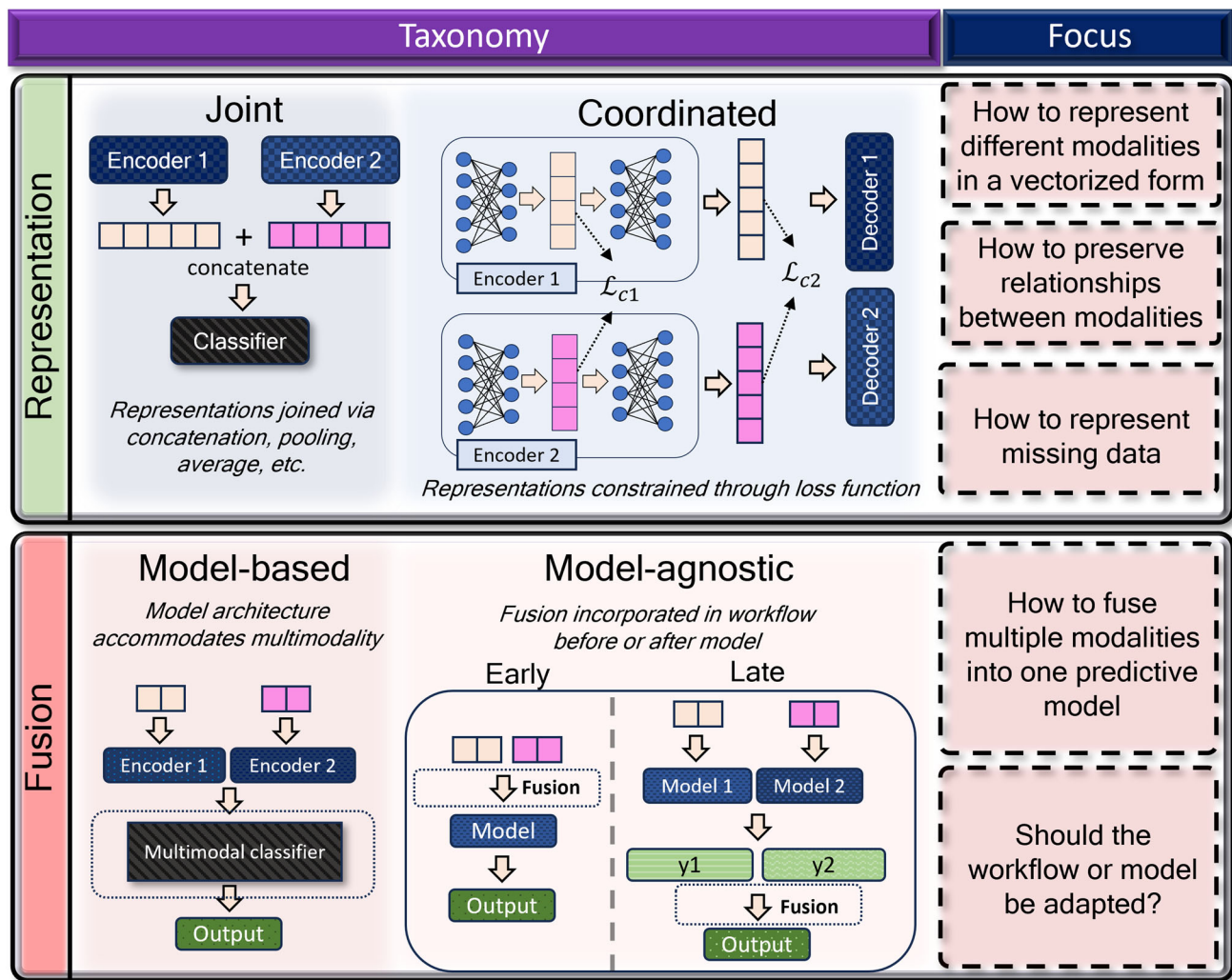
Fig. 2 A graphical representation of the taxonomical sublevels of multimodal representation and fusion, and the focus of each challenge. Multimodal representation can be categorized into whether the representations are joined into a single vector (*joint*) or separately constructed to be influenced by each other (*coordinated*). Multimodal fusion can be distinguished by whether a model is uniquely constructed to fuse the modalities (*model-based*), or whether fusion occurs before or after the model step (*model-agnostic*)

Another approach is use of autoencoders, which condense image representations into lower-dimensional context vectors that can be decoded to reconstruct the original image. *Multimodal autoencoders* have been applied to MRI modalities in Hamghalam et al. (2021) and in this example were also utilized to impute representations for missing modalities.

Another approach for multimodal representation could be through the use of *disentanglement networks*, which can separate latent properties of an image into separate vectors. In such cases, an image is given as input and the autoencoder is often split in such a way that two vectors are produced as intermediate pathways, where joining the intermediate vectors should result in the original input. Each intermediate pathway is often constrained by a separate loss function term to encourage separation of each pathway into the desired latent characteristics. In this way, one input image can be represented by two separate vectors, each representing a disjointed characteristic of the image. This disentanglement method has been applied in Jiang and Veeraraghavan (2020) to separate context in CT and MRI from their style so that one modality can be converted in to the other. It was also applied for a single modality in Bône et al. (2020) to separate "shape" and "appearance" representations of an input, which could potentially be applied to different imaging modalities to extract only similar shapes from each.

When two or more vectorized modalities are combined into a model, they are typically combined in one of two ways: (1) joint, or (2) coordinated representations. A *joint representation* is characterized by aggregation of the vectors at some point in the process, whereby vector representations from two
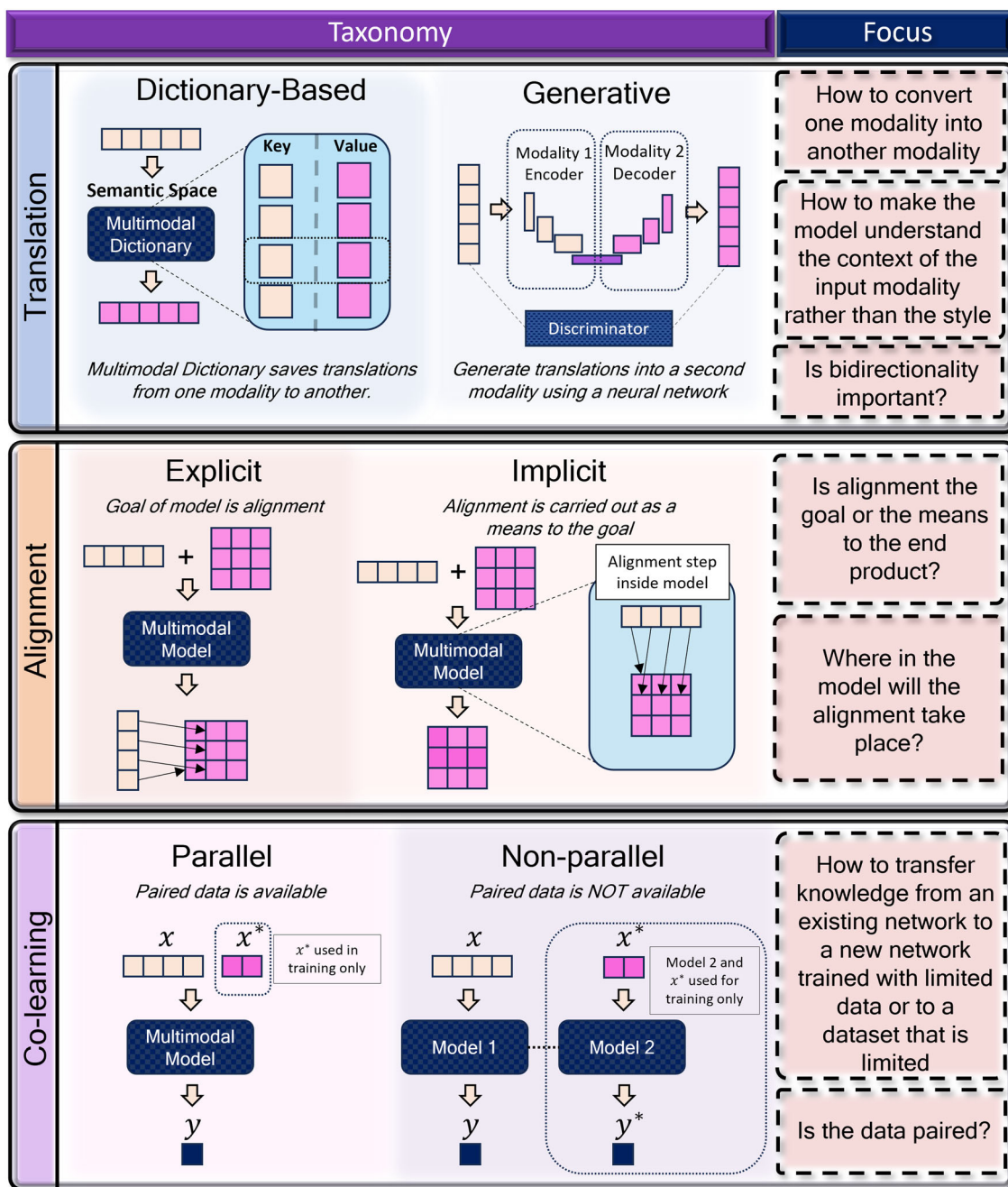
Fig. 3 A graphical representation of the taxonomical sublevels of multimodal translation, alignment and co-learning, and the focus of each challenge. In *translation*, models are distinguished based on whether they require use of a dictionary to save associations between modalities (*dictionary-based*), or if the associations are learned in a multimodal network (*generative*). In *alignment*, distinction is made depending on the *purpose* of the alignment, whether as the goal (*explicit*) or as an intermediate step towards the goal output (*implicit*). In *co-learning*, a distinction is made between the use of *parallel* (paired) multimodal data, or *non-parallel* (unpaired) multimodal data. In co-learning models, one of the modalities is only used in training but does not appear in testing

Table 1 Literature relating to the five challenges of multimodal machine learning by the datatype analyzed

| Challenge | Datatype | | | |
	MRI	CT	PET	EHR
Representation	Hamghalam et al. (2021) and Zhang et al. (2022)	Daza et al. (2020) and Zhou et al. (2023)		Daza et al. (2020), Sonsbeek and Worring (2020), Zhang et al. (2023) and Wang et al. (2023)
Fusion	Azcona et al. (2020), Neubauer et al. (2020), Zhou et al. (2020), Wang et al. (25021), Zhang et al. (2022), Rudie et al. (2022), Liu et al. (2023), Zhang et al. (2023) and Zhou et al. (2023)	Daza et al. (2020), Neubauer et al. (2020), Yang et al. (2020), Bhalodia et al. (2021) and Carbonell et al. (2023)	Neubauer et al. (2020)	Daza et al. (2020), Yang et al. (2020), Sonsbeek and Worring (2020), Zhou et al. (2020), Vivar et al. (2020), Li et al. (2021), Bhalodia et al. (2021), Cui et al. (2022) and Khosravi et al. (2022)
Translation	Jiang and Veeraraghavan (2020), Hu et al. (2020), Guo et al. (2020), Shin et al. (2020) and Takagi and Nishimoto (2023)	Jiang and Veeraraghavan (2020) and Zhu et al. (2020)	Hu et al. (2020) and Shin et al. (2020)	
Alignment	Wang et al. (25021) and Zhang et al. (2022)	Yang et al. (2020), Leroy et al. (2023), Zhou et al. (2023) and Li et al. (2023)		Yang et al. (2020) and Li et al. (2023)
Co-learning	Varsavsky et al. (2020), Yang et al. (2020), Bui et al. (2020), Hu et al. (2021), Pei et al. (2023), Jafari et al. (2022), Xue et al. (2020) and Dong et al. (2022)	Xue et al. (2020), Hu et al. (2021), Pei et al. (2023), Jafari et al. (2022) and Dong et al. (2022)		Xing et al. (2022)

| Challenge | Datatype | | | | |
	Hist	Ultrasound	Genomic	X-ray	Fundus
Representation				Sonsbeek and Worring (2020)	Zhou et al. (2023)
Fusion	Chen et al. (2020), Li et al. (2021) and Cui et al. (2022)	Habib et al. (2020)	Chen et al. (2020) and Cui et al. (2022)	Sonsbeek and Worring (2020), Habib et al. (2020), Cui et al. (2022), Carbonell et al. (2023), Khosravi et al. (2022)	Zhou et al. (2023)
Translation					
Alignment	Leroy et al. (2023)				
Co-learning	Xiong et al. (2020)		Chauhan et al. (2020)		Zhou et al. (2023)

separate modalities are joined together into a single vector form through methods such as aggregation, concatenation or summation. Joint representation is both a common and effective strategy for representation; however, a joint strategy such as concatenation is often constricted to serving in situations where both modalities are available at train- and test-time (one exception using Boltzmann Machines can be found in Srivastava and Salakhutdinov (2014)). If a modality has the potential to be missing, a joint strategy such as aggregation via weighted means could be a better option (Li et al., 2021; Chen et al., 2020; Zhou et al., 2023; Cui et al., 2022). Using mathematical notation from Baltrusaitis et al. (2019), we can denote joint representations x_m as the following:

$$x_m = f(x_1, ..., x_n) \tag{1}$$

This denotes that feature vectors x_i, $i = 1...n$ are combined in some way through a function f to create a new representation space x_m. By the contrary, *coordinated representations* are represented as the following:

$$f(x_1) \sim g(x_2), \tag{2}$$

whereby a function designed to create representations for one modality may be constrained (represented by $\sim$) by a similar function from another modality, with the assumption that relationships between data points in the first modality should be relatively well-preserved in the second modality.

Joint representations tend to be the most common approach to representing two or more modalities together in a model because it is perhaps the most straightforward approach. For example, joining vectorized multimodal data together through concatenation before entering a model tends to be one of the most direct approaches to joint representation. Sonsbeek and Worring (2020), for example, chest x-rays are combined with text data from electronic health records in a vectorized form using a pretrained model first. Then, the vectors from each modality are sent individually through two attention-based blocks, then concatenated into a joint feature space to predict a possible cardiovascular disease and generate a free-text "impression" of the condition. Other joint representation models follow simpler methods, simply extracting baseline features from a pretrained model and concatenating them Daza et al. (2020); Yang et al. (2020).

Although coordinated representations have traditionally tended to be more challenging to implement, the convenience of neural network architectural and loss adjustments have resulted in increased traction in publications embodying coordinated representations (Xing et al., 2022; Wang et al., 2023; Chauhan et al., 2020; Radford et al., 2021; Zhang et al., 2022; Bhalodia et al., 2021). One of the most notable of these in recent AI approaches is OpenAI's Contrastive Language-Image Pre-Training (CLIP) model, which forms representations for OpenAI's DALL·E 2 (Radford et al., 2021; Ramesh et al., 2022) and uses a contrastive-learning approach to shape both image embeddings of entire images to match text embeddings of entire captions describing those images. The representations learned from CLIP were demonstrated to not only perform well in zero-shot image-to-text or text-to-image models, but also to produce representations that could outpace baseline supervised learning methods. In a biomedical context, similar models abound, including ConVIRT, a predecessor and forerunner for CLIP (Zhang et al., 2022), and related works (Bhalodia et al., 2021).

Coordinated approaches are especially useful in co-learning. Chauhan et al. (2020), which employs a subset of co-learning called privileged information, the geometric forms of each modality are not joined into a single vector representation. Instead, network weights are encouraged to produce similar output vectors for each modality and ultimately the same classifications. This constraint warps the space of chest x-ray representations closer to the space of text representations, with the assumption that this coordinated strategy provides chest x-ray representations more useful information because of the text modality. For more on privileged information, see the Sect. 2.5 below.

In the biomedical sphere, where models are built to prioritize biologically- or clinically-relevant outcomes, quality of representations may often be overlooked or overshadowed by emphasis on optimization of prediction accuracy. However, there is conceptual value in building good multimodal representations. If models are constructed to ensure that similar concepts in different modalities also demonstrate cross-modal similarity, then there is greater confidence that an accurate model is understanding cross-modal relationships. While building good cross-modal representations for indexing images on the Internet like in the CLIP model is a digestible challenge, building similar cross-modal representations for medical data presents a far more formidable challenge due to data paucity. OpenAI's proprietary WebTextImage dataset, used for CLIP, contains 400 million examples, a sample size that is as of yet unheard of for any kind of biomedical imaging data. Until such a dataset is released, bioinformaticians must often rely on the ability to leverage pretrained models and transfer learning strategies for optimal results amidst low resources to leverage big data for good representations on smaller data.

2.2 Fusion

Next, we discuss challenges in multimodal fusion. This topic is a natural segue from the discussion of representation because many multimodal representations are subsequently fed into a discriminatory model. Multimodal fusion entails the utilization of methods to combine representations from more than one modality into a classification, regression,

or segmentation model. According to Baltrusaitis et al. (2019), fusion models can be classified into two subcategories: model-agnostic and model-based approaches. The term *"model-agnostic"* refers to methods for multimodal fusion occurring either before or after the model execution and typically does not involve altering the prediction model itself. Model-agnostic approaches can further be delineated by the stage at which the fusion of modalities occurs, either early in the model (prior to output generation) or late in the model (such as ensemble models, where outputs from multiple models are combined). Additionally, hybrid models, incorporating a blend of both early and late fusion, have been proposed (Carbonell et al., 2023). In contrast, a *model-based* approach entails special adjustments to the predictive model to ensure it handles each modality uniquely.

While model-agnostic methods remain pertinent as useful strategies for multimodal fusion, the overwhelming popularity of neural networks has led to a predominant shift towards model-based methods in recent years. These model-based methods involve innovative loss functions and architectures designed to handle each modality differently. One common model-based fusion strategy is multimodal *multiple instance learning (MIL)*, where multiple context vectors for each modality are generated and subsequently aggregated into a single representation leading to the output classification. The method for aggregation varies across studies, with attention-based approaches, emphasizing specific characteristics of each modality, being a common choice (Li et al., 2021; Chen et al., 2020; Zhou et al., 2023; Cui et al., 2022).

The construction of a good model architecture is crucial; however, challenges associated with fusion are often highly contextual, and thus it is important to understand what kinds of data are being utilized in recent models and what problems they try to solve. Most multimodal models understandably incorporate MRI modalities, given that MR images are a natural multimodal medium. Consequently, studies incorporating MRI such as Azcona et al. (2020), which aims to classify Alzheimer's Disease severity, and Zhou et al. (2020), predicting overall survival in brain tumor patients, exemplify the type of research often prevalent in multimodal image-based clinical application publications. Brain-based ML studies are also popular because of the wide availability of brain images and a strong interest in applying ML models in clinical neuroradiology. However, recent models encompass a myriad of other clinical scenarios predicting lung cancer presence (Daza et al., 2020), segmenting soft tissue sarcomas (Neubauer et al., 2020), classifying breast lesions (Habib et al., 2020), and predicting therapy response (Yang et al., 2020), among others, by amalgamating and cross-referencing modalities such as CT images (Daza et al., 2020; Neubauer et al., 2020), blood tests (Yang et al., 2020), electronic health record (EHR) data (Yang et al., 2020; Sonsbeek and Worring, 2020; Daza et al., 2020), mammography

images (Habib et al., 2020), and ultrasound (Habib et al., 2020).

Multimodal fusion models are emerging as the gold standard for clinical-assisted interventions due to the recognition that diagnosis and prognosis in real-world clinical settings are often multimodal problems. However, these models are not without limitations. For one, standardization across equipment manufacturers or measurement protocols can affect model performance dramatically, and this issue becomes more pronounced as more modalities are incorporated into a model. Second, while fusion models attempt to mimic real-world clinical practice, they face practical challenges that can limit their utility. For instance, physicians may face various roadblocks to obtaining all model input variables due to a lack of permission from insurance companies to perform all needed tests or time constraints. These issues underscore challenges associated with missing modalities, and several studies have attempted to address this concern (Carbonell et al., 2023; Zhang et al., 2022; Cui et al., 2022; Wang et al., 2023; Liu et al., 2023). However, incorporating mechanisms to account for missing modalities in a model is not yet a common practice for most multimodal biomedical models.

Lastly, many models are not configured to make predictions that adapt with additional variables. Most models necessitate all variables to be present at the time of operation, meaning that, even if all tests are conducted, the model can only make a decision once all test results have been obtained. In conclusion, in the dynamic and fast-paced environment of hospitals and other care centers, even accurate models may not be suitable for practical use, unless also coupled with mechanisms to handle missing data.

2.3 Translation

In multimodal translation, a model is devised to operate as a mapping entity facilitating the transformation from one modality to another. This involves the conversion of input contextual data, such as CT scans, into an alternative contextual data format, such as MRI scans. Before the rise of modern *generative* methods leveraging multimodal generative adversarial networks (GANs) or diffusion models to generate one modality from another, translation via *dictionary-based* methods was common, which typically involved a bimodal dictionary whereby a single entry would contain a key belonging to one modality and a corresponding value belonging to the other modality. Dictionary-based translation was uncommon in biomedical research but popular in NLP fields as a way to convert images into text and vice versa (Liao et al., 2022; Reed et al., 2016). The current ascendancy of generative models and the availability of associated coding packages have since catalyzed the growth of innovative translational studies applying generative approaches.

Presently, generative models encompass a broad spectrum of potential applications both within and beyond the biomedical domain. Outside the medical sphere, generative models find utility in NLP settings, particularly in text-to-image models like DALL·E 2 and Midjourney (Liao et al., 2022; Ramesh et al., 2022; Oppenlaender, 2022). Additionally, they are employed in style transfer and other aesthetic computer vision techniques (Huang et al., 2021; Cao et al., 2018; Zhu et al., 2017; Liu et al., 2018; Palsson et al., 2018; Zhang and Wang, 2020). Within the biomedical realm, generative models have proven efficacious in creating virtual stains for unstained histopathological tissues which would typically undergo hemotoxylin/eosin staining (Lu et al., 2021). Furthermore, these models serve as prominent tools for sample generation (Tseng et al., 2017; Piacentino et al., 2021; Choi et al., 2017), particularly in scenarios with limited sample sizes (Chen et al., 2021). Despite the potential diversity of multimodal translation involving any two modalities, predominant translational efforts in the biomedical realm currently revolve around mapping one imaging modality to another, a paradigm recognized as image-to-image translation.

In the contemporary landscape, the integration of simplistic generative models into a clinical context are declining in visibility, while methods employing specialized architectures tailored to the involved modalities are acknowledged for advancing the state-of-the-art in translational work. Within this context, two notable generative translation paradigms for biomedicine are explored: (1) medical image generation models, and (2) segmentation mask models. In the former, many studies attempt to form models that are bidirectional, whereby the intended output can be placed back as input and return an image similar to the original input image. Bui et al. (2020), this is resolved by generating deformation fields that map changes in the T1-weighted sequence modality of MRI to the T2-weighted sequence modality. Hu et al. (2020), separate forward and backward training processes are defined whereby an encoder representing PET images is utilized to understand the underlying distribution of that modality, allowing for more realistic synthetic generated images from MRI. In one unidirectional example, Shin et al. (2020) modifies a pix2pix conditional GAN network to allow Alzheimer's disease classification to influence synthetic PET image generation. In another interesting example, Takagi and Nishimoto (2023) use functional MRI (fMRI) scans and diffusion models to attempt to recreate images of what their subjects had seen. Similarly, diffusion models and magnetoencephalography (MEG) are utilized by Meta for real-time prediction from brain activity of what patients had seen visually (Benchetrit et al., 2023).

In the second potential application, image segmentation models in multimodal image-to-image translation must handle additional challenges, creating both a way to generate the output modality as well as a way to segment it. Jiang and Veeraraghavan (2020), a generative model converts CT to MRI segmentation. In a reverse problem to image segmentation, Guo et al. (2020) attempts to synthesize multimodal MRI examples of lesions with only a binary lesion mask and a multimodal MRI Atlas. In this study, six CNN-based discriminators are utilized to ensure the authentic appearance of background, brain, and lesion, respectively, in synthesized images.

Multimodal translation still remains an exciting but formidable challenge. In NLP and beyond, there have been remarkable successes observed in new image generation within text-to-image models beyond the biomedical sphere. However, the adoption of translation models in biomedical work is evolving at a more measured pace, with applications extending beyond demonstrative feasibility to practical utility remaining limited. Arguments in favor of biomedical translation models are predominantly centered around sample generation for datasets with limited sizes, as the generated medical images must adhere to stringent accuracy requirements. Similar to other challenges in multimodal research, translation models would greatly benefit from training on more expansive and diverse datasets. However, with the increasing digitization of medical records and a refined understanding of de-identification protocols and data sharing rights, the evolution of this field holds considerable promise.

2.4 Alignment

Multimodal alignment involves aligning two related modalities, often in either a spatial or temporal way. Multimodal alignment can be conducted either *explicitly* as a direct end goal, or *implicitly*, as a means to the end goal, which could be translation or classification of an input. One example of *explicit alignment* in a biomedical context is image registration. Leroy et al. (2023) highlights one approach to multimodal image registration, where histopathology slides are aligned to their (x, y, z) coordinates in a three-dimensional CT volume. Another is in Chen et al. (2023), where surgical video was aligned to a text description of what is happening in the video. On the other hand, an example of multimodal *implicit alignment* could be the temporal alignment of multiple clinical tests to understand a patients progress over time. Such an analysis was conducted in Yang et al. (2020), where the authors built a customized multi-layer perceptron (MLP) called SimTA to predict response to therapy intervention at a future time step based on results from previous tests and interventions.

Literature surrounding alignment has increased since the rise of attention-based models in 2016. The concept of "attention," which relates to aligning representations in a way that is contextually relevant, is a unimodal alignment paradigm with origins in machine translation and NLP (Bahdanau et al., 2015). An example use of attention in NLP could be mod-

els which try to learn, based on order and word choice of an input sentence, where the subject of the sentence is so that the response can address the input topic. In imaging, attention can be used to highlight important parts of an image that are most likely to contribute to a class prediction. Vaswani et al. (2017), introduced a more sophisticated attention network, named transformers, an encoder-decoder-style architecture based on repeated projection heads where attention learning takes place. Transformers and attention were originally applied to natural language (Vaswani et al., 2017; Bahdanau et al., 2015; Devlin et al., 2019) but have since been applied to images (Parmar et al., 2018; Dosovitskiy et al., 2021), including histopathology slides (Lu et al., 2021; Chen et al., 2020) and protein prediction (Tunyasuvunakool et al., 2021). *Multimodal transformers* were introduced in 2019, also developed for the natural language community (Tsai et al., 2019). While these multimodal transformers do not contain the same encoder-decoder structure of a traditional transformer architecture, they are hallmarked by crossmodal attention heads, where one modality's sequences intermingle with another modality's sequences.

Although typical transformers themselves are not multimodal, they often constitute in multimodal models. The SimTA network mentioned above borrowed the positional encoding property of transformers to align multimodal inputs in time to predict therapy response (Yang et al., 2020). Many models taking advantage of visual transformers (ViT) have also utilized pretrained transformers trained on images for multimodal fusion models. In both the TransBTS (Wang et al., 25021) and mmFormer models (Zhang et al., 2022), a transformer is utilized on a vector composed of an amalgamation of information from multiple modalities of MRI, which may imply that the transformer attention heads here are aligning information from multiple modalities represented via aggregate latent vectors. The ultimate function of transformers is a form of implicit alignment, and it can be assumed here that this alignment is multimodal.

Transformer models have brought a new and largely successful approach to alignment, sparking widespread interest in their applications in biomedical use. Transformers for NLP have also engendered new interest in Large Language Models (LLMs), which are already being applied to biomedical contexts (Tinn et al., 2023) and probing new questions about its potential use as a knowledge base for biomedical questions (Sung et al., 2021).

2.5 Co-learning

In this last section exploring recent research in multimodal machine learning, the area of co-learning is examined, a field which has recently garnered a strong momentum in both unimodal and multimodal domains. In multimodal co-learning, knowledge learned from one modality is often used to assist learning of a second modality. This first modality which transfers knowledge is often leveraged only at train-time but is not required at test-time. Co-learning is classified in Baltrusaitis et al. (2019) as either parallel or non-parallel. In *parallel* co-learning, paired samples of modalities which share the same instance are fed into a co-learning model. By contrast, in *non-parallel* co-learning, both modalities are included in a model but are not required to be paired.

While co-learning can embody a variety of topics such as conceptual grounding and zero-shot learning, this work focuses on the use of transfer learning in biomedicine. In *multimodal transfer learning*, a model trained on a higher quality or more plentiful modality is employed to assist in the training of a model designed for a second modality which is often noisier or smaller in sample size. Transfer learning can be conducted in both parallel and non-parallel paradigms. This work focuses on one parallel form of transfer learning called privileged learning, and one non-parallel form of transfer learning called domain adaptation. A visual representation of these approaches be seen in Fig. 4.

2.5.1 Privileged Learning

Privileged learning originates from the mathematician Vladmir Vapnik and his ideas of knowledge transfer with the support vector machine for privileged learning (SVM+) model (Vapnik and Vashist, 2009). The concept of privileged learning introduces the idea that predictions for a low-signal, low-cost modality can be assisted by incorporating a high-signal, high-cost modality (privileged information) in training only, while at test-time only the low-cost modality is needed. Vapnik and Vashist (2009), Vapnik illustrates this concept through the analogy of a teacher (privileged information) distilling knowledge to a student (low-cost modality) before the student takes a test. Although a useful concept, the field is relatively under-explored compared to other areas of co-learning. One challenge to applying privileged learning models was that Vapnik's SVM+ model was one of few available before the widespread use of neural networks. Furthermore, it demands that the modality deemed "privileged" must confer high accuracy on its own in order to ensure that its contribution to the model is positive. Since then, neural networks have encouraged newer renditions of privileged information models that allow more flexibility of use (Lambert et al., 2018; Shaikh et al., 2020; Sabeti et al., 2021).

Recently, privileged learning has emerged as a growing subset of biomedical literature, and understandably so. Many multimodal models today require health care professionals to gather a slew of patient information and are not trained to handle missing data. Therefore, the ability to minimize the number of required input data while still utilizing the predictive power of multiple modalities can be useful in real-world clinical settings. Hu et al. (2020) for example,
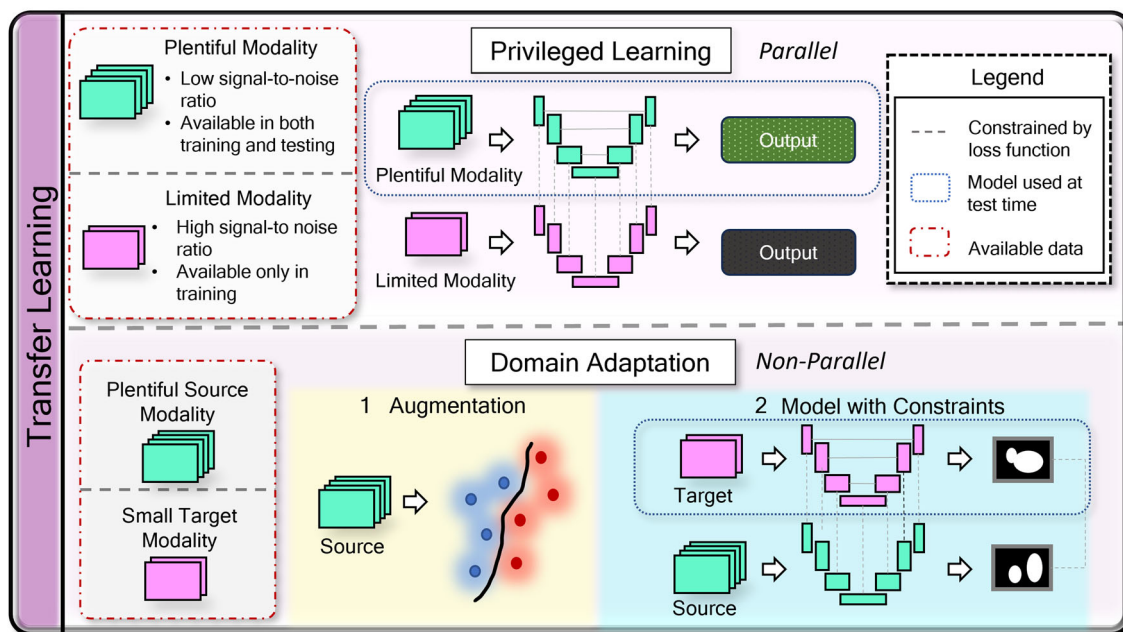
Fig. 4 Two types of transfer learning described in this work are privileged learning (top) and domain adaptation (bottom). In privileged learning, a plentiful set consisting of data which is normally of low cost but also low signal-to-noise ratio is available in both training and testing, while a limited gold-standard quality set is used for training only. In this example, the plentiful set is used to train the target model, while the limited set constrains the model parameters to increase the model's ability to associate the low-cost modality with the ground truth. In domain adaptation, there is a target dataset which consists of a few samples and a source dataset consisting of plenty of samples. If the target data is too small to build a reliable model in training, source data can be augmented to make the model more robust. Else, the target model could be trained with few examples, while a second source model is used to help make the target model more generalizable

the authors attempt to train a segmentation network where at train-time the "teacher network" contains four MR image modalities, but at test-time the "student network" contains only T1-weighted images, the standard modality used in preoperative neurosurgery and radiology. Chauhan et al. (2020), chest x-rays and written text from their respective radiology reports are used to train a model where only chest x-rays are available at test-time.

In privileged models based on traditional approaches (before deep neural networks), privileged information can be embedded in the model either through an alteration of allowable error ("slack variables" from SVM+) (Vapnik and Vashist, 2009), or through decision trees constructed with non-privileged features to mimic the discriminative ability of privileged features (Random Forest+) (Warner et al., 2022; Moradi et al., 2016). In a deep learning model, privileged learning is often achieved through the use of additional loss functions which attempt to constrain latent and output vectors from the non-privileged modality to mimic those from the combined privileged and non-privileged models (Hu et al., 2020; Xing et al., 2022). For example, in Chauhan et al. (2020), encoders for each modality are compared and cross entropy loss is calculated for each modality separately. The sum of these allows the chest x-ray network to freely train for only the chest x-ray modality while being constrained

through the overall loss function to borrow encoding methods from the text network, which also strives to build an accurate model.

While privileged learning models can be applied where data is missing, users should heed caution when applying models in situations where there is systematic bias in reporting. Those who train privileged models without considering subject matter may inadvertently be choosing to include all their complete data in training and their incomplete data in testing. However, in clinical scenarios, data are often incomplete because a patient either did not qualify for a test (perhaps their condition was seen as not "dire enough" to warrant a test) or their situation was too serious to require a test (for example, a patient in septic shock may not pause to undergo a chest x-ray because they are in the middle of a medical emergency). Therefore, while applying data to highly complex models is a common approach in computer science, the context of the data and potential underlying biases need to be considered first to ensure a practical and well-developed model.

2.5.2 Domain Adaptation

Domain adaptation has been shown to be useful in biomedical data science applications where a provided dataset may

be too small or costly to utilize for more advanced methods such as deep learning, but where a somewhat similar (albeit larger) dataset can be trained by such methods. The smaller dataset for which we want to train the model is called the "target" dataset and the larger dataset which will be used to assist the model with the learning task and provide better contextualization is called the "source" dataset. Domain adaptation strategies are often tailored to single modalities such as camera imaging or MRI, where measurements of an observed variable differ based on an instrument's post-processing techniques or acquisition parameters (Xiong et al., 2020; Varsavsky et al., 2020; Yang et al., 2020). However, the distinct characteristics arising from disparate instruments or acquisition settings can lead to considerable shifts in data distribution and feature representations, mirroring the challenges faced in true multimodal contexts. Therefore, the discussion of uni-modal domain adaptation is a relevant starting point for multimodal domain adaptation, as it covers approaches to mitigate significant deviations within data that may seem similar but are represented differently. Additionally, understanding how to mitigate the impact of such variations helps one to understand ways to construct multimodal machine learning systems that confront similar challenges. We also discuss relevant multimodal domain adaptation approaches in biomedicine, which have typically consisted of applying CT images as a source domain to train an MRI target model or vice versa (Chiou et al., 2020; Xue et al., 2020; Pei et al., 2023; Jafari et al., 2022; Dong et al., 2022).

One way to train a model to adapt to different domains is through augmentation of the input data, which "generalizes" the model to interpret outside of the domain of the original data. Xiong et al. (2020), a data augmentation framework for fundus images in diabetic retinopathy (DR) is proposed to offset the domain differences of utilizing different cameras. The authors show that subtracting local average color, blurring, adaptive local contrast enhancement, and a specialized principal component analysis (PCA) strategy can increase both R^2 values for age prediction and DR classification area under the receiver operating curve (AUROC) on test sets where either some domain information is known *a priori* and also where no information is known, respectively. In another method which attempts to augment the source domain into more examples in the target style, Chiou et al. (2020) split the source image into latent content and style vectors, using the content vectors in a style-transfer model reminiscent of cycleGAN to feed as examples with the target domain into a segmentation network (Zhu et al., 2017). In other applications, data augmentation for domain generalization may be executed utilizing simpler affine transformations (Varsavsky et al., 2020). This demonstrates the utility of data augmentation strategies in more broadly defining decision boundaries where target domains differ from the source.

A second strategy for domain adaptation involves constraining neural network functions trained on a target domain by creating loss functions which require alignment with a source domain model. Varsavsky et al. (2020), a framework for adapting segmentation models at test-time is proposed, whereby an adversarial loss trains a target-based U-Net to be as similar to a source-based U-Net as possible. Then a paired-consistency loss with adversarial examples is utilized to fine-tune the decision boundary to include morphologically similar data points. In a specifically multimodal segmentation-based model, Xue et al. (2020) attempts to create two side-by-side networks, a segmenter and an edge generator, which both encourage the source and target output to be as similar as possible to each other. In the final loss function, the edge generator is used to constrain the segmenter in such a way as to promote better edge consistency in the target domain. In yet another, simpler example, domain adaptation to a target domain is performed in Hu et al. (2021) by taking a network trained on the source domain and simply adjusting the parameters of the batch normalization layer.

Domain adaptation in biomedicine can be a common problem where instrument models or parameters change. Among multimodal co-learning methods, most networks are constructed as segmentation networks for MRI and CT because they are similar imaging domains, although measuring different things. While CT carries distinct meaning in its pixels (measured in Hounsfield Units), MRI pixel intensities are not standardized and usually require normalization, which could pose challenges to this multimodal problem. Additionally, MRI carries much more detail than CT scans, which necessitates the model to understand contextual boundaries of objects much more than a unimodal case with only CT or MRI.

3 Discussion

The rapidly evolving landscape of artificial intelligence (AI) both within the biomedical field and beyond has posed a substantial challenge in composing this survey. Our aim is to provide the reader with a comprehensive overview of the challenges and contemporary approaches to multimodal machine learning in image-based, clinically relevant biomedicine. However, it is essential to acknowledge that our endeavor cannot be fully comprehensive due to the dynamic nature of the field and the sheer volume of emerging literature within the biomedical domain and its periphery. This robust growth has led to a race among industry and research institutions to integrate the latest cutting-edge models into the healthcare sector, with a particular emphasis on the introduction of "large language models" (LLMs). In recent years, there has been an emergence of market-level insights into the future of healthcare and machine learning, as exemplified by

the incorporation of machine learning models into wearable devices such as the Apple Watch and Fitbit devices for the detection of atrial fibrillation (Perino et al., 2021; Lubitz et al., 2022). This begs the question: *where does this transformative journey lead us?*

Healthcare professionals and physicians already embrace the concept of multimodal cognitive models in their diagnostic and prognostic practices, signaling that such computer models based on multimodal frameworks are likely to endure within the biomedical landscape. However, for these models to be effectively integrated into clinical settings, they must exhibit flexibility that aligns with the clinical environment. If the ultimate goal is to seamlessly incorporate these AI advancements into clinical practice, a fundamental question arises: how can these models be practically implemented on-site? Presently, most available software tools for clinicians are intended as auxiliary aids, but healthcare professionals have voiced concerns regarding the potential for increased computational workload, alert fatigue, and the limitations imposed by Electronic Health Record (EHR) interfaces (Ruiter et al., 2015; Ancker et al., 2017). Therefore, it is paramount to ensure that any additional software introduced into clinical settings serves as an asset rather than a hindrance.

Another pertinent issue emerging from these discussions pertains to the dynamics between clinical decision support systems (CDSS) and healthcare providers. What occurs when a computer-generated recommendation contradicts a physician's judgment? This dilemma is not new, as evidenced by a classic case recounted by Evans et al. (1998), where physicians were granted the choice to either follow or disregard a CDSS for antibiotic prescription. Intriguingly, the group provided with the choice exhibited suboptimal performance compared to both the physician-only and computer-only groups. Consequently, it is unsurprising that some healthcare professionals maintain a cautious approach to computer decision support systems (Adamson and Welch, 2019; Silcox et al., 2020). Questions arise regarding the accountability of physicians if they ignore a correct computer-generated decision and the responsibility of software developers if a physician follows an erroneous computer-generated recommendation.

A pivotal ingredient notably under-represented in many CDSS models, which could help alleviate discrepancies between computer-generated and human decisions, is the incorporation of uncertainty quantification, grounded calibration, interpretability and explainability. These factors have been discussed in previous literature, underscoring the critical role of explainability in ensuring the long-term success of CDSS-related endeavors (Reddy, 2022; Khosravi et al., 2022; Kwon et al., 2020; Abdar et al., 2021).

The domain of multimodal machine learning for medically oriented image-based clinical support has garnered increasing attention in recent years. This interest has been stimulated by advances in computer science architecture and computing hardware, the availability of vast and publicly accessible data, innovative model architectures tailored for limited datasets, and the growing demand for applications in clinical and biomedical contexts. Recent studies have showcased the ability to generate synthetic images in one modality based on another (as outlined in Sect. 2.3), align multiple modalities (Sect. 2.4), and transfer latent features from one modality to train another (Sect. 2.5), among other advancements. These developments offer a promising outlook for a field that is still relatively new. However, it is also imperative to remain vigilant regarding the prevention of data biases and under-representation in ML models to maximize the potential of these technologies.

Despite these promising developments, the field faces significant hurdles, notably the lack of readily available "big data" in the medical domain. For instance, the routine digitization of histopathology slides remains a challenging goal in many healthcare facilities. Data sharing among medical institutions is fraught with challenges around appropriate procedures for the responsible sharing of patient data under institutional, national and international patient privacy regulations.

Advancing the field will likely entail overcoming these hurdles, ensuring more extensive sharing of de-identified data from research publications and greater participation in establishment of standardized public repositories for data. Dissemination of both code and pretrained model weights would also enable greater knowledge-sharing and repeatability. Models that incorporate uncertainty quantification, explainability, and strategies to account for missing data are particularly advantageous. For more guidance on building appropriate multimodal AI models in healthcare, one can refer to the World Health Organization's new ethics and governance guidelines for large multimodal models (World Health Organization, 2024).

In conclusion, the field of multimodal machine learning in biomedicine has experienced rapid growth in each of its challenge areas of representation, fusion, translation, alignment, and co-learning. Given the recent advancements in deep learning models, escalating interest in multimodality, and the necessity for multimodal applications in healthcare, it is likely that the field will continue to mature and broaden its clinical applications. In this ever-evolving intersection of AI and healthcare, the imperative for responsible innovation resonates strongly. The future of multimodal machine learning in the biomedical sphere presents immense potential but also mandates a dedication to ethical principles encompassing data privacy, accountability, and transparent collaboration between human professionals and AI systems. As we navigate this transformative journey, the collective effort, ethical stewardship, and adherence to best practices will ensure the

realization of the benefits of AI and multimodal machine learning, making healthcare more efficient, accurate, and accessible, all while safeguarding the well-being of patients and upholding the procedural and ethical standards of clinical practice.

Data Availability No data outside of those referenced has been used in this survey. Key papers have been summarized in Table 1.

Declarations

Conflict of Interest The Authors declare no competing financial interests but the following competing non-financial interests: A.R. serves as a member for Voxel Analytics, LLC. C.E.K.'s institution receives salary support for service as editor of Radiology: Artificial Intelligence.

References

Abdar, M., Samami, M., Mahmoodabad, S. D., Doan, T., Mazoure, B., Hashemifesharaki, R., Liu, L., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning. *Computers in Biology and Medicine, 135*, 104418. https://doi.org/10.1016/j.compbiomed.2021.104418

Adamson, A. S., & Welch, H. G. (2019). Machine learning and the cancer-diagnosis problem—No gold standard. *New England Journal of Medicine, 381*(24), 2285–2287. https://doi.org/10.1056/nejmp1907407

Ancker, J. S., Edwards, A., Nosal, S., Hauser, D., Mauer, E., & Kaushal, R. (2017). Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Medical Informatics and Decision Making*. https://doi.org/10.1186/s12911-017-0430-8

Azcona, E. A., Besson, P., Wu, Y., Punjabi, A., Martersteck, A., Dravid, A., Parrish, T. B., Bandt, S. K., & Katsaggelos, A. K. (2020). Interpretation of brain morphology in association to Alzheimer's disease dementia classification using graph convolutional networks on triangulated meshes. In *Shape in medical imaging* (pp. 95–107). Springer. https://doi.org/10.1007/978-3-030-61056-2_8

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Y. Bengio, Y. LeCun (Eds.), *3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference track proceedings*. arxiv:1409.0473.

Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 423–443. https://doi.org/10.1109/tpami.2018.2798607

Benchetrit, Y., Banville, H., & King, J.-R. (2023). Brain decoding: Toward real-time reconstruction of visual perception. arXiv:2310.19812

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798–1828. https://doi.org/10.1109/tpami.2013.50

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y., & Ramesh, A. (2023). Improving image generation with better captions. https://cdn.openai.com/papers/dall-e-3.pdf.

Bhalodia, R., Hatamizadeh, A., Tam, L., Xu, Z., Wang, X., Turkbey, E., & Xu, D. (2021). Improving pneumonia localization via cross-attention on medical images and reports. In *Medical image computing and computer assisted intervention—MICCAI 2021* (pp. 571–581). Springer. https://doi.org/10.1007/978-3-030-87196-3_53

Bône, A., Vernhet, P., Colliot, O., & Durrleman, S. (2020). Learning joint shape and appearance representations with metamorphic auto-encoders. In *Medical image computing and computer assisted intervention—MICCAI 2020* (pp. 202–211). Springer. https://doi.org/10.1007/978-3-030-59710-8_20

Bui, T. D., Nguyen, M., Le, N., & Luu, K. (2020). Flow-based deformation guidance for unpaired multi-contrast MRI image-to-image translation. In *Medical image computing and computer assisted intervention—MICCAI 2020* (pp. 728–737). Springer. https://doi.org/10.1007/978-3-030-59713-9_70

Cao, K., Liao, J., & Yuan, L. (2018). CariGANs. *ACM Transactions on Graphics, 37*(6), 1–14. https://doi.org/10.1145/3272127.3275046

Carbonell, E.L., Shen, Y., Yang, X., & Ke, J. (2023). COVID-19 pneumonia classification with transformer from incomplete modalities. In *Lecture notes in computer science* (pp. 379–388). Springer. https://doi.org/10.1007/978-3-031-43904-9_37

Chauhan, G., Liao, R., Wells, W., Andreas, J., Wang, X., Berkowitz, S., Horng, S., Szolovits, P., & Golland, P. (2020). Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In *Medical image computing and computer assisted intervention—MICCAI 2020* (pp. 529–539). Springer. https://doi.org/10.1007/978-3-030-59713-9_51

Chen, Z., Guo, Q., Yeung, L. K. T., Chan, D. T. M., Lei, Z., Liu, H., & Wang, J. (2023). Surgical video captioning with mutual-modal concept alignment. In *Lecture notes in computer science* (pp. 24–34). Springer. https://doi.org/10.1007/978-3-031-43996-4_3

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering, 5*(6), 493–497. https://doi.org/10.1038/s41551-021-00751-8

Chen, R. J., Lu, M. Y., Wang, J., Williamson, D. F. K., Rodig, S. J., Lindeman, N. I., & Mahmood, F. (2020). Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*. https://doi.org/10.1109/tmi.2020.3021387

Chiou, E., Giganti, F., Punwani, S., Kokkinos, I., & Panagiotaki, E. (2020). Harnessing uncertainty in domain adaptation for MRI prostate lesion segmentation. In *Medical image computing and*

computer assisted intervention—MICCAI 2020 (pp. 510–520). Springer. https://doi.org/10.1007/978-3-030-59710-8_50

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017). Generating multi-label discrete patient records using generative adversarial networks. In F. Doshi-Velez, J. Fackler, D. Kale, R. Ranganath, B. Wallace, J. Wiens (Eds.), *Proceedings of the 2nd machine learning for healthcare conference. Proceedings of machine learning research* (Vol. 68, pp. 286–305). PMLR. https://proceedings.mlr.press/v68/choi17a.html

Cui, C., Liu, H., Liu, Q., Deng, R., Asad, Z., Wang, Y., Zhao, S., Yang, H., Landman, B. A., & Huo, Y. (2022). Survival prediction of brain cancer with incomplete radiology, pathology, genomic, and demographic data. In *Lecture notes in computer science* (pp. 626–635). Springer. https://doi.org/10.1007/978-3-031-16443-9_60

Daza, L., Castillo, A., Escobar, M., Valencia, S., Pinzón, B., & Arbeláez, P. (2020). LUCAS: LUng CAncer screening with multimodal biomarkers. In *Multimodal learning for clinical decision support and clinical image-based procedures* (pp. 115–124). Springer. https://doi.org/10.1007/978-3-030-60946-7_12

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *North American chapter of the association for computational linguistics*. https://aclanthology.org/N19-1423.pdf

Dong, D., Fu, G., Li, J., Pei, Y., & Chen, Y. (2022). An unsupervised domain adaptation brain CT segmentation method across image modalities and diseases. *Expert Systems with Applications, 207*, 118016. https://doi.org/10.1016/j.eswa.2022.118016

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In *International conference on learning representations*. arxiv:2010.11929

Evans, R. S., Pestotnik, S. L., Classen, D. C., Clemmer, T. P., Weaver, L. K., Orme, J. F., Lloyd, J. F., & Burke, J. P. (1998). A computer-assisted management program for antibiotics and other antiinfective agents. *New England Journal of Medicine, 338*(4), 232–238. https://doi.org/10.1056/nejm199801223380406

Guo, P., Wang, P., Zhou, J., Patel, V.M., & Jiang, S. (2020). Lesion mask-based simultaneous synthesis of anatomic and molecular MR images using a GAN. In *Medical image computing and computer assisted intervention—MICCAI 2020* (pp. 104–113). Springer. https://doi.org/10.1007/978-3-030-59713-9_11

Habib, G., Kiryati, N., Sklair-Levy, M., Shalmon, A., Neiman, O. H., Weidenfeld, R. F., Yagil, Y., Konen, E., & Mayer, A. (2020). Automatic breast lesion classification by joint neural analysis of mammography and ultrasound. In *Multimodal learning for clinical decision support and clinical image-based procedures* (pp. 125–135). Springer. https://doi.org/10.1007/978-3-030-60946-7_13

Hamghalam, M., Frangi, A.F., Lei, B., & Simpson, A. L. (2021). Modality completion via gaussian process prior variational autoencoders for multi-modal glioma segmentation. In: *Medical image computing and computer assisted intervention—MICCAI 2021* (pp. 442–452). Springer. https://doi.org/10.1007/978-3-030-87234-2_42

Hu, M., Maillard, M., Zhang, Y., Ciceri, T., Barbera, G. L., Bloch, I., & Gori, P. (2020). Knowledge distillation from multi-modal to mono-modal segmentation networks. In *Medical image computing and computer assisted intervention—MICCAI 2020* (pp. 772–781). Springer. https://doi.org/10.1007/978-3-030-59710-8_75

Hu, S., Shen, Y., Wang, S., & Lei, B. (2020). Brain MR to PET synthesis via bidirectional generative adversarial network. In *Medical image computing and computer assisted intervention—MICCAI 2020* (pp. 698–707). Springer. https://doi.org/10.1007/978-3-030-59713-9_67

Hu, M., Song, T., Gu, Y., Luo, X., Chen, J., Chen, Y., Zhang, Y., & Zhang, S. (2021). Fully test-time adaptation for image segmentation. In *Medical image computing and computer assisted intervention—MICCAI 2021* (pp. 251–260). Springer. https://doi.org/10.1007/978-3-030-87199-4_24

Huang, Z., Chen, S., Zhang, J., & Shan, H. (2021). PFA-GAN: Progressive face aging with generative adversarial network. *IEEE Transactions on Information Forensics and Security, 16*, 2031–2045. https://doi.org/10.1109/tifs.2020.3047753

Jafari, M., Francis, S., Garibaldi, J. M., & Chen, X. (2022). LMISA: A lightweight multi-modality image segmentation network via domain adaptation using gradient magnitude and shape constraint. *Medical Image Analysis, 81*, 102536. https://doi.org/10.1016/j.media.2022.102536

Jiang, J., & Veeraraghavan, H. (2020). Unified cross-modality feature disentangler for unsupervised multi-domain MRI abdomen organs segmentation. In *Medical image computing and computer assisted intervention—MICCAI 2020* (pp. 347–358). Springer. https://doi.org/10.1007/978-3-030-59713-9_34

Khosravi, B., Rouzrokh, P., Kremers, H. M., Larson, D. R., Johnson, Q. J., Faghani, S., Kremers, W. K., Erickson, B. J., Sierra, R. J., Taunton, M. J., & Wyles, C. C. (2022). Patient-specific hip arthroplasty dislocation risk calculator: An explainable multimodal machine learning–based approach. *Radiology: Artificial Intelligence*. https://doi.org/10.1148/ryai.220067

Kwon, Y., Won, J.-H., Kim, B. J., & Paik, M. C. (2020). Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis, 142*, 106816. https://doi.org/10.1016/j.csda.2019.106816

Lambert, J., Sener, O., & Savarese, S. (2018). Deep learning under privileged information using heteroscedastic dropout. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2018/papers/Lambert_Deep_Learning_Under_CVPR_2018_paper.pdf

Leroy, A., Cafaro, A., Gessain, G., Champagnac, A., Grégoire, V., Deutsch, E., Lepetit, V., & Paragios, N. (2023). StructuRegNet: Structure-guided multimodal 2D-3D registration. In *Lecture notes in computer science* (pp. 771–780). Springer. https://doi.org/10.1007/978-3-031-43999-5_73

Li, T. Z., Still, J. M., Xu, K., Lee, H. H., Cai, L. Y., Krishnan, A. R., Gao, R., Khan, M. S., Antic, S., Kammer, M., Sandler, K. L., Maldonado, F., Landman, B. A., & Lasko, T. A. (2023) Longitudinal multimodal transformer integrating imaging and latent clinical signatures from routine EHRs for pulmonary nodule classification. In *Lecture notes in computer science* (pp. 649–659). Springer. https://doi.org/10.1007/978-3-031-43895-0_61

Li, H., Yang, F., Xing, X., Zhao, Y., Zhang, J., Liu, Y., Han, M., Huang, J., Wang, L., & Yao, J. (2021). Multi-modal multi-instance learning using weakly correlated histopathological images and tabular clinical information. In *Medical image computing and computer assisted intervention—MICCAI 2021* (pp. 529–539). Springer. https://doi.org/10.1007/978-3-030-87237-3_51

Liao, W., Hu, K., Yang, M. Y., & Rosenhahn, B. (2022). Text to image generation with semantic-spatial aware GAN. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr52688.2022.01765

Liu, H., Michelini, P.N., & Zhu, D. (2018). Artsy-GAN: A style transfer system with improved quality, diversity and performance. In *2018 24th international conference on pattern recognition (ICPR)*. IEEE. https://doi.org/10.1109/icpr.2018.8546172

Liu, Z., Wei, J., Li, R., & Zhou, J. (2023). SFusion: Self-attention based n-to-one multimodal fusion block. In *Lecture notes in computer science* (pp. 159–169). Springer. https://doi.org/10.1007/978-3-031-43895-0_15

Lubitz, S. A., Faranesh, A. Z., Selvaggi, C., Atlas, S. J., McManus, D. D., Singer, D. E., Pagoto, S., McConnell, M. V., Pantelopoulos, A., & Foulkes, A. S. (2022). Detection of atrial fibrillation in a large population using wearable devices: The Fitbit heart study. *Circula-*

tion, 146(19), 1415–1424. https://doi.org/10.1161/circulationaha.122.060291

Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering, 5*(6), 555–570. https://doi.org/10.1038/s41551-020-00682-w

Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering, 5*(6), 555–570. https://doi.org/10.1038/s41551-020-00682-w

Metz, R. (2022). AI won an art contest, and artists are furious. Warner Bros. Discovery. https://www.cnn.com/2022/09/03/tech/ai-art-fair-winner-controversy/index.html

Moradi, M., Syeda-Mahmood, T., & Hor, S. (2016). Tree-based transforms for privileged learning. In *Machine learning in medical imaging* (pp. 188–195). Springer. https://doi.org/10.1007/978-3-319-47157-0_23

Neubauer, T., Wimmer, M., Berg, A., Major, D., Lenis, D., Beyer, T., Saponjski, J., & Bühler, K. (2020). Soft tissue sarcoma co-segmentation in combined MRI and PET/CT data. In *Multimodal learning for clinical decision support and clinical image-based procedures* (pp. 97–105). Springer. https://doi.org/10.1007/978-3-030-60946-7_10

OpenAI: GPT-4 Technical Report. (2023). arXiv:2303.08774.

Oppenlaender, J. (2022). The creativity of text-to-image generation. In *Proceedings of the 25th international academic mindtrek conference*. ACM. https://doi.org/10.1145/3569219.3569352

Palsson, S., Agustsson, E., Timofte, R., & Van Gool, L. (2018). Generative adversarial style transfer networks for face aging. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) workshops*. https://openaccess.thecvf.com/content_cvpr_2018_workshops/papers/w41/Palsson_Generative_Adversarial_Style_CVPR_2018_paper.pdf

Parmar, N. J., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., & Tran, D. (2018). Image transformer. In *International conference on machine learning (ICML)*. http://proceedings.mlr.press/v80/parmar18a.html

Pei, C., Wu, F., Yang, M., Pan, L., Ding, W., Dong, J., Huang, L., & Zhuang, X. (2023). Multi-source domain adaptation for medical image segmentation. *IEEE Transactions on Medical Imaging*. https://doi.org/10.1109/tmi.2023.3346285

Perino, A. C., Gummidipundi, S. E., Lee, J., Hedlin, H., Garcia, A., Ferris, T., Balasubramanian, V., Gardner, R.M., Cheung, L., Hung, G., Granger, C. B., Kowey, P., Rumsfeld, J. S., Russo, A. M., True Hills, M., Talati, N., Nag, D., Tsay, D., Desai, S., Desai, M., Mahaffey, K. W., Turakhia, M. P., & Perez, M. V. (2021). Arrhythmias other than atrial fibrillation in those with an irregular pulse detected with a smartwatch: Findings from the Apple heart study. *Circulation: Arrhythmia and Electrophysiology*. https://doi.org/10.1161/circep.121.010063

Piacentino, E., Guarner, A., & Angulo, C. (2021). Generating synthetic ECGs using GANs for anonymizing healthcare data. *Electronics, 10*(4), 389. https://doi.org/10.3390/electronics10040389

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning. Proceedings of machine learning research* (Vol. 139, pp. 8748–8763). PMLR. https://proceedings.mlr.press/v139/radford21a.html

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. arXiv:2204.06125.

Reddy, S. (2022). Explainability and artificial intelligence in medicine. *The Lancet Digital Health, 4*(4), 214–215. https://doi.org/10.1016/s2589-7500(22)00029-2

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. In: M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd international conference on machine learning. Proceedings of machine learning research* (Vol. 48, pp. 1060–1069). PMLR. https://proceedings.mlr.press/v48/reed16.html

Rudie, J. D., Calabrese, E., Saluja, R., Weiss, D., Colby, J. B., Cha, S., Hess, C. P., Rauschecker, A. M., Sugrue, L. P., & Villanueva-Meyer, J. E. (2022). Longitudinal assessment of posttreatment diffuse glioma tissue volumes with three-dimensional convolutional neural networks. *Radiology: Artificial Intelligence*. https://doi.org/10.1148/ryai.210243

Ruiter, H., Liaschenko, J., & Angus, J. (2015). Problems with the electronic health record. *Nursing Philosophy, 17*(1), 49–58. https://doi.org/10.1111/nup.12112

Sabeti, E., Drews, J., Reamaroon, N., Warner, E., Sjoding, M. W., Gryak, J., & Najarian, K. (2021). Learning using partially available privileged information and label uncertainty: Application in detection of acute respiratory distress syndrome. *IEEE Journal of Biomedical and Health Informatics, 25*(3), 784–796. https://doi.org/10.1109/jbhi.2020.3008601

Shaikh, T. A., Ali, R., & Beg, M. M. S. (2020). Transfer learning privileged information fuels CAD diagnosis of breast cancer. *Machine Vision and Applications*. https://doi.org/10.1007/s00138-020-01058-5

Shin, H.-C., , Ihsani, A., Xu, Z., Mandava, S., Sreenivas, S. T., Forster, C., & Cha, J. (2020). GANDALF: Generative adversarial networks with discriminator-adaptive loss fine-tuning for Alzheimer's disease diagnosis from MRI. In *Medical image computing and computer assisted intervention—MICCAI 2020* (pp. 688–697). Springer. https://doi.org/10.1007/978-3-030-59713-9_66

Silcox, C., Dentzer, S., & Bates, D. W. (2020). AI-enabled clinical decision support software: A "trust and value checklist" for clinicians. *NEJM Catalyst*. https://doi.org/10.1056/cat.20.0212

Sonsbeek, T., & Worring, M. (2020). Towards automated diagnosis with attentive multi-modal learning using electronic health records and chest X-rays. In *Multimodal learning for clinical decision support and clinical image-based procedures* (pp. 106–114). Springer. https://doi.org/10.1007/978-3-030-60946-7_11

Srivastava, N., & Salakhutdinov, R. (2014). Multimodal learning with deep Boltzmann machines. *Journal of Machine Learning Research, 15*(84), 2949–2980.

Sung, M., Lee, J., Yi, S. S., Jeon, M., Kim, S., & Kang, J. (2021). Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 4723–4734). Association for Computational Linguistics. ACL. arXiv:2109.07154. https://aclanthology.org/2021.emnlp-main.388.pdf

Takagi, Y., & Nishimoto, S. (2023). High-resolution image reconstruction with latent diffusion models from human brain activity. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr52729.2023.01389

Tinn, R., Cheng, H., Gu, Y., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2023). Fine-tuning large neural language models for biomedical natural language processing. *Patterns, 4*(4), 100729. https://doi.org/10.1016/j.patter.2023.100729

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1656

Tseng, H.-H., Luo, Y., Cui, S., Chien, J.-T., Haken, R. K. T., & Naqa, I. E. (2017). Deep reinforcement learning for automated radiation adaptation in lung cancer. *Medical Physics, 44*(12), 6690–6705. https://doi.org/10.1002/mp.12625

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., & Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*. https://doi.org/10.1038/s41586-021-03828-1

Vapnik, V., & Vashist, A. (2009). A new learning paradigm: Learning using privileged information. *Neural Networks, 22*(5–6), 544–557. https://doi.org/10.1016/j.neunet.2009.06.042

Varsavsky, T., Orbes-Arteaga, M., Sudre, C. H., Graham, M. S., Nachev, P., & Cardoso, M. J. (2020). Test-time unsupervised domain adaptation. In *Medical image computing and computer assisted intervention—MICCAI 2020* (pp. 428–436). Springer. https://doi.org/10.1007/978-3-030-59710-8_42

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. CoRR arXiv:1706.03762

Vivar, G., Mullakaeva, K., Zwergal, A., Navab, N., & Ahmadi, S.-A. (2020). Peri-diagnostic decision support through cost-efficient feature acquisition at test-time. In *Medical image computing and computer assisted intervention—MICCAI 2020* (pp. 572–581). Springer. https://doi.org/10.1007/978-3-030-59713-9_55

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., & Li, J. (2021). Trans-BTS: Multimodal brain tumor segmentation using transformer. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021* (pp. 109–119). Springer. https://doi.org/10.1007/978-3-030-87193-2_11

Wang, H., Ma, C., Zhang, J., Zhang, Y., Avery, J., Hull, L., & Carneiro, G. (2023). Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In *Lecture notes in computer science* (pp. 216–226). Springer. https://doi.org/10.1007/978-3-031-43901-8_21

Warner, E., Al-Turkestani, N., Bianchi, J., Gurgel, M. L., Cevidanes, L., & Rao, A. (2022). Predicting osteoarthritis of the temporomandibular joint using random forest with privileged information. In *Ethical and philosophical issues in medical imaging, multimodal learning and fusion across scales for clinical decision support, and topological data analysis for biomedical imaging* (pp. 77–86). Springer. https://doi.org/10.1007/978-3-031-23223-7_7

World Health Organization. (2024). Ethics and governance of artificial intelligence for health: Guidance on large multimodal models, Geneva. https://www.who.int/publications/i/item/9789240084759

Xing, X., Chen, Z., Zhu, M., Hou, Y., Gao, Z., & Yuan, Y. (2022). Discrepancy and gradient-guided multi-modal knowledge distillation for pathological glioma grading. In *Lecture notes in computer science* (pp. 636–646. Springer. https://doi.org/10.1007/978-3-031-16443-9_61

Xiong, J., He, A. W., Fu, M., Hu, X., Zhang, Y., Liu, C., Zhao, X., & Ge, Z. (2020). Improve unseen domain generalization via enhanced local color transformation. In *Medical image computing and computer assisted intervention—MICCAI 2020* (pp. 433–443). Springer. https://doi.org/10.1007/978-3-030-59713-9_42

Xue, Y., Feng, S., Zhang, Y., Zhang, X., & Wang, Y. (2020). Dual-task self-supervision for cross-modality domain adaptation. In *Medical image computing and computer assisted intervention—MICCAI 2020* (pp. 408–417). Springer. https://doi.org/10.1007/978-3-030-59710-8_40

Yang, J., Chen, J., Kuang, K., Lin, T., He, J., & Ni, B. (2020). MIA-prognosis: A deep learning framework to predict therapy response. In *Medical image computing and computer assisted intervention—MICCAI 2020* (pp. 211–220). Springer. https://doi.org/10.1007/978-3-030-59713-9_21

Yang, Y., Wang, N., Yang, H., Sun, J., & Xu, Z. (2020). Model-driven deep attention network for ultra-fast compressive sensing MRI guided by cross-contrast MR image. In *Medical image computing and computer assisted intervention—MICCAI 2020* (pp. 188–198). Springer. https://doi.org/10.1007/978-3-030-59713-9_19

Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., Zhang, Y., He, Z., & Zheng, Y. (2022). mmFormer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *Lecture notes in computer science* (pp. 107–117). Springer. https://doi.org/10.1007/978-3-031-16443-9_11

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., & Langlotz, C.P. (2022). Contrastive learning of medical visual representations from paired images and text. In *Proceedings of machine learning research* (Vol. 182, pp. 1–24). Machine Learning for Healthcare, PMLR. https://proceedings.mlr.press/v182/zhang22a/zhang22a.pdf

Zhang, L., Na, S., Liu, T., Zhu, D., & Huang, J. (2023). Multimodal deep fusion in hyperbolic space for mild cognitive impairment study. In *Lecture notes in computer science* (pp. 674–684). Springer. https://doi.org/10.1007/978-3-031-43904-9_65

Zhang, F., & Wang, C. (2020). MSGAN: Generative adversarial networks for image seasonal style transfer. *IEEE Access, 8*, 104830–104840. https://doi.org/10.1109/access.2020.2999750

Zhou, T., Fu, H., Zhang, Y., Zhang, C., Lu, X., Shen, J., & Shao, L. (2020). M2net: Multi-modal multi-channel network for overall survival time prediction of brain tumor patients. In *Medical image computing and computer assisted intervention—MICCAI 2020* (pp. 221–231). Springer. https://doi.org/10.1007/978-3-030-59713-9_22

Zhou, Y., Yang, G., Zhou, Y., Ding, D., & Zhao, J. (2023). Representation, alignment, fusion: A generic transformer-based framework for multi-modal glaucoma recognition. In *Lecture notes in computer science* (pp. 704–713). Springer. https://doi.org/10.1007/978-3-031-43990-2_66

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv:1703.10593.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE international conference on computer vision (ICCV)*. IEEE. https://doi.org/10.1109/iccv.2017.244

Zhu, Y., Tang, Y., Tang, Y., Elton, D. C., Lee, S., Pickhardt, P. J., & Summers, R. M. (2020). Cross-domain medical image translation by shared latent gaussian mixture model. In *Medical image computing and computer assisted intervention—MICCAI 2020* (pp. 379–389). Springer. https://doi.org/10.1007/978-3-030-59713-9_37